Perception, Cognition, and Effectiveness of Visualizations with Applications in Science and Engineering

Perception, Cognition, and Effectiveness of Visualizations with Applications in Science and Engineering

A dissertation presented

by

Michelle A. Borkin

to

The School of Engineering & Applied Sciences

in partial fulfillment of the requirements

for the degree of

Doctor of Philosophy

in the subject of

Applied Physics

Harvard University

Cambridge, Massachusetts

April 2014

Prof. Hanspeter Pfister & Prof. Alyssa A. Goodman Michelle A. Borkin

Perception, Cognition, and Effectiveness of Visualizations with Applications in Science and Engineering

Abstract

Visualization is a powerful tool for data exploration and analysis. With data ever-increasing in quantity and becoming integrated into our daily lives, having effective visualizations is necessary. But how does one design an effective visualization? To answer this question we need to understand how humans perceive, process, and understand visualizations. Through visualization evaluation studies we can gain deeper insight into the basic perception and cognition theory of visualizations, both through domain-specific case studies as well as generalized laboratory experiments.

This dissertation presents the results of four evaluation studies, each of which contributes new knowledge to the theory of perception and cognition of visualizations. The results of these studies include a deeper clearer understanding of how color, data representation dimensionality, spatial layout, and visual complexity affect a visualization's effectiveness, as well as how visualization types and visual attributes affect the memorability of a visualization.

We first present the results of two domain-specific case study evaluations. The first study is in the field of biomedicine in which we developed a new heart disease diagnostic tool, and conducted a study to evaluate the effectiveness of 2D versus 3D data representations as well as color maps. In the second study, we developed a new visualization tool for filesystem provenance data with applications in computer science and the sciences more broadly. We additionally developed a new time-based hierarchical node grouping method. We then conducted a study to evaluate the effectiveness of the new tool with its radial layout versus the conventional node-link diagram, and the new node grouping method. Finally, we discuss the results of two generalized studies designed to understand what makes a visualization memorable. In the first evaluation we focused on visualization memorability and conducted an online study using Amazon's Mechanical Turk with hundreds of users and thousands of visualizations. For the second evaluation we designed an eye-tracking laboratory study to gain insight into precisely which elements of a visualization contribute to memorability as well as visualization recognition and recall.

Contents

List of Figures

List of Tables

Dedicated to Mom and Dad,

for their endless support, encouragement, and love.

And to Moli, my best friend and the love of my life.

Author List

The following authors contributed to this thesis:

- Chapter 3: Krzysztof Z. Gajos, Amanda Peters, Dimitrios Mitsouras, Simone Melchionna, Frank J. Rybicki, Charles L. Feldman, and Hanspeter Pfister.
- Chapter 4: Chelsea S. Yeh, Madelaine Boyd, Peter Macko, Krzysztof Z. Gajos, Margo Seltzer, and Hanspeter Pfister.
- Chapter 5: Azalea A. Vo, Zoya Bylinskii, Phillip Isola, Shashank Sunkavalli, Aude Oliva, and Hanspeter Pfister.
- Chapter 6: Zoya Bylinskii, Constance May Bainbridge, Chelsea S. Yeh, Daniel Borkin, Hanspeter Pfister, and Aude Oliva.

Acknowledgments

I am deeply grateful for the guidance and support of my advisors, Professors Hanspeter Pfister and Alyssa Goodman, under whose mentorship I was introduced to the field of visualization. It has been an immense pleasure to work with Hanspeter. I will forever be thankful for his belief in my academic and professional potential, and for his motivation throughout my graduate studies. I am also indebted for the support and inspiration Alyssa has provided not just throughout my Ph.D. but also in the work which led me to my dissertation over the past many years. She saw potential in me as a high school student, and has helped encourage me ever since to reach for the stars.

My thanks goes out to my many wonderful collaborators from whom I have learned so much including Krzysztof Gajos, Aude Oliva, Margo Seltzer, Charles Feldman, Efthimios Kaxiras, Frank Rybicki, Zoya Bylinskii, and Prof. Eric Mazur who supported me as a member of my thesis committee. They have all made this dissertation possible, and contributed to my personal growth as a researcher. I also thank Madelaine Boyd, Azalea Vo, Chelsea Yeh, Sean Yeh, and Daniel Borkin who all inspired me while I was serving as their research advisor.

Thanks to the current and past members of the Visual Computing Group (VCG) at SEAS with whom I have been honored to be a group member and had as my academic family throughout my graduate studies. I also greatly appreciate all those who supported me at SEAS including the Student Affairs Office.

My research would not have been possible without the financial support of my graduate fellowships. I thank my advisors for their help and for their encouragement to apply for the fellowships. This dissertation work was supported by the National Defense Science & Engineering Graduate Fellowship (NDSEG) Program and the National Science Foundation Graduate Research Fellowship Program (NSF GRFP).

I would also like to acknowledge and thank a few very special people who inspired me and shepherded me along my journey, but are sadly not with us to see my dissertation completed: Prof. John Huchra, who fostered my passion for astronomy and physics as a young novice researcher, Dr. Nelson Bernstein, who believed in me and opened my eyes to the world of biology and medicine, and Dr. Michael Adlerstein (Ph.D. 1971, Applied Physics, Harvard University) whose passion for physics and math inspired me to pursue my love of science all the way to his alma mater.

I am grateful to my entire wonderful family for their love and support including my grandmother Marilyn Borkin, my geologist uncle Chris Ruisaard for nurturing my interest in science, and my newest family Dr. Yi Luo, Qing Gao, and Douglass Luo. I want to especially thank my parents, Dr. Sheldon and Sandy Borkin. They have been patient and encouraging, and have always believed in me. Their love and support has been instrumental to my success. I would also in particular like to thank my Dad for sharing with me his infectious love of science, and introducing me to it at a young age through building model rockets and dinosaurs, teaching me the constellations on nighttime walks, and sharing weekly Saturday morning basement laboratory experiments. I would not be where I am today without his infinite supply of passion, genius, and wisdom. Finally, I want to thank my husband Moli Luo for his unconditional love and support through the final years of my dissertation. I could not imagine getting through this journey of completing my dissertation without Moli, nor any of life's future journeys.

1

Introduction

We live in a world where data is ubiquitous. We interact with, need, and use data on a daily basis. This can be as pedestrian as checking the weather, the daily stock prices, or map to a destination in your car. This can be as specialized as a doctor viewing radiology images of a patient or a physicist exploring data output from a particle accelerator. In this world we need methods for exploring and understanding all forms of data, with our machines helping people to extract the greatest knowledge from their data. Visualization is one such powerful method.

Visualization is the static or interactive visual representation of data to reinforce human cognition. Common visualization techniques include plots or graphs, tables, diagrams, or multidimensional renderings. The data visualized can either be spatial (e.g., geographic, anatomical, etc.) or non-spatial (e.g., quantitative values, calculations, etc.). The human visual system accounts for almost 50% of processing in the brain [127]. Data visualization harnesses this processing power and can enable visual data exploration and discovery.

How does one design an *effective* visualization? To answer and address this question, we need to understand how humans see and process a visualization. For a given data set, there may be many ways to represent the data and encode multiple variables. However certain methods will inherently be more effective than others for a given task and data type due to underlying human perceptual and cognitive properties.

Through careful experimentation and evaluation, we can both better understand existing theories of perception and cognition as applied to visualization and also develop and lay the groundwork for a new basic understanding of these principles in visualization. This theory of visualization perception and cognition can be derived and evaluated in general through one of two methods. The first method is evaluation of perception and cognition within the context of a visualization case study in a specific domain. These case study applications typically have visualization challenges and visual encoding methodology whose solutions have broader or more generalizable implications, both in terms of the visual encoding but also the perceptual and cognitive properties. The other method is evaluation and experiments to develop basic theory not specific to one domain application but with fundamentally broad applicable and generalizable results as the main goal of the study. In this thesis both methods are utilized in order to more thoroughly understand and validate existing theories as well as develop fundamentally new theories of visualization cognition and perception.

Through both specific case study evaluations in science and engineering topics as well as general experiments, this thesis contributes new fundamental knowledge on what types and features of visualizations affect a person's perception and cognition of a visualization as well as make them more effective for specific domain tasks. The results of these studies and experiments include, as outlined in detail below, a deeper and clearer understanding of how color, data representation dimensionality, layout, and visual complexity affect a visualization's effectiveness, as well as how visualization types and visual attributes affect the memorability of a visualization.

1.1 Summary of Original Contributions

To address this need for a more through and solid understanding of visualization perception and cognition for the design of effective visualizations, this thesis focuses on both specific case studies as well as general experiments to evaluate and create fundamental theory. The major contributions of this thesis are outlined below:

Evaluation of Artery Visualizations for Heart Disease Diagnosis: Heart disease is the number one killer in the United States, and finding indicators of the disease at an early stage is critical for treatment and prevention. In this chapter we evaluate visualization techniques that enable the diagnosis of coronary artery disease. A key physical quantity of medical interest is endothelial shear stress (ESS).

Low ESS has been associated with sites of lesion formation and rapid progression of disease in the coronary arteries. Having effective visualizations of a patient's ESS data is vital for the quick and thorough non-invasive evaluation by a cardiologist. We present a task taxonomy for hemodynamics based on a formative user study with domain experts. Based on the results of this study we developed HemoVis, an interactive visualization application for heart disease diagnosis that uses a novel 2D tree diagram representation of coronary artery trees. We present the results of a formal quantitative user study with domain experts that evaluates the effect of 2D versus 3D artery representations and of color maps on identifying regions of low ESS. We show statistically significant results demonstrating that our 2D visualizations are more accurate and efficient than 3D representations, and that a perceptually appropriate color map leads to fewer diagnostic mistakes than a rainbow color map.

Evaluation of Filesystem Provenance Visualization Tools: Having effective visualizations of filesystem provenance data is valuable for understanding its complex hierarchical structure. The most common visual representation of provenance data is the node-link diagram. While effective for understanding local activity, the node-link diagram fails to offer a high-level summary of activity and inter-relationships within the data. We present a new tool, InProv, which displays filesystem provenance with an interactive radial-based tree layout. The tool also utilizes a new time-based hierarchical node grouping method for filesystem provenance data we developed to match the user's mental model and make data exploration more intuitive. We compared InProv to a conventional node-link based tool, Orbiter, in a quantitative evaluation with real users of filesystem provenance data including provenance data experts, IT professionals, and computational scientists.

We also compared in the evaluation our new node grouping method to a conventional method. The results demonstrate that InProv results in higher accuracy in identifying system activity than Orbiter with large complex data sets. The results also show that our new time-based hierarchical node grouping method improves performance in both tools, and participants found both tools significantly easier to use with the new time-based node grouping method. Subjective measures show that participants found InProv to require less mental activity, less physical activity, less work, and is less stressful to use. Our study also reveals one of the first cases of gender differences in visualization; both genders had comparable performance with InProv, but women had a significantly lower average accuracy (56%) compared to men (70%) with Orbiter.

What Makes a Visualization Memorable?: An ongoing debate in the visualization community concerns the role that visualization types play in data understanding. In human cognition, understanding and memorability are intertwined. As a first step towards being able to ask questions about impact and effectiveness, here we ask: "What makes a visualization memorable?" We ran the largest scale visualization study to date using 2,070 single-panel visualizations, categorized with visualization type (e.g., bar chart, line graph, etc.), collected from news media sites, government reports, scientific journals, and infographic sources. Each visualization was annotated with additional attributes, including ratings for data-ink ratios and visual densities. Using Amazon's Mechanical Turk, we collected memorability scores for hundreds of these visualizations, and discovered that observers are consistent in which visualizations they find memorable and forgettable. We find intuitive results (e.g., attributes like color and the inclusion of a human recognizable object enhance memorability) and less intuitive results (e.g., common graphs are less memorable than unique visualization types). Altogether our findings suggest that quantifying memorability is a general metric of the utility of information, an essential step towards determining how to design effective visualizations.

Eye-tracking Study for Visualization Recognition and Recall: What do you remember about a visualization? In this study we designed an eye-tracking lab experiment to evaluate which features of a visualization (e.g., data, title, text, etc.) affect its recognition and recall. After exploring hundreds of visualizations for 10 seconds each (*encoding* phase), observers performed a *recognition* task ("Have you seen this visualization before?") followed by a *recall* task ("Can you describe what you remember about this visualization?"). Analyses of eye movements reveal that whether visualizations are recognized or not trigger different eye movement patterns between the encoding and recognition phases, suggesting that the way a person explores a visualization is a marker of his or her memory. Based on the results of this experiment, we present guidelines on what types and features of visualizations will affect a viewer's recognizability and recall of a visualization.

1.2 Thesis Structure

This thesis is divided into seven chapters. In Chapter 2, an overview of previous related work and concepts is discussed. Chapters 3 and 4 are examples of case study evaluations: Chapter 3 addresses a case study evaluation in cardiology for artery visualizations, and Chapter 4 addresses a case study evaluation in filesystem provenance graph visualizations. Chapters 5 and 6 are examples of experiments for the development of fundamental theory: Chapter 5 presents an experiment to study the memorability of a visualization, and Chapter 6 presents an experiment to study the recognizability and recall of a visualization. Finally, in Chapter 7 the thesis concludes with a summary of the contributions and discussion of directions for future work.

2

Related Work

This chapter presents research related to the overarching topics discussed in this thesis. We start with a discussion of perception and cognition theory as applied to visualization, followed by a discussion of evaluation techniques and design methodology. Finally, we summarize some of the key trends, challenges, and techniques for data visualization in the physical sciences and engineering to frame the topics discussed in Chapters 3 and 4.

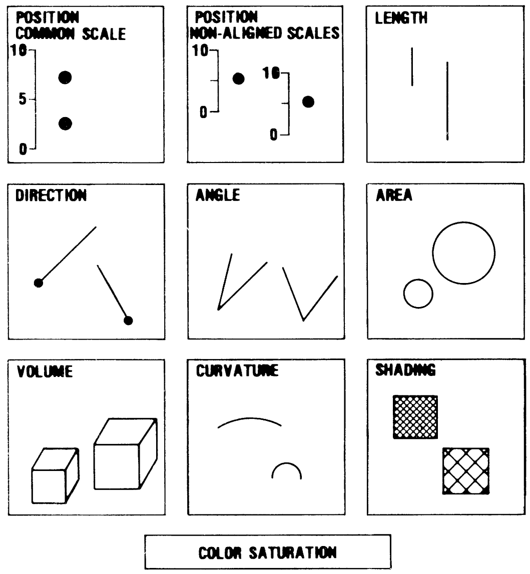

Figure 2.1: Elementary perceptual tasks in order from most to least accurate for quantitative value judgement: positions along a common scale, positions along nonaligned scales, length, direction, angle, area, volume, curvature, shading, and color saturation. (Figure from [37].)

2.1 Perception & Cognition Theory in Visualization

Understanding fundamental human perceptual and cognitive properties and how they affect a visualization's interpretation is a very important and on-going area of active research. This topic spans many disciplines and has roots in the fields of psychology, cognitive sciences, statistics, and the computer science fields of human-computer interaction (HCI) and visualization.

In the visualization community, many important works have studied how different visualization types are perceived, and the effect of different data types and tasks [18, 37, 65, 92, 111, 132, 183, 184]. For example, in Cleveland & McGill [37], different visual encodings for data were evaluated in a lab study experiment in order to better understand whether humans can more accurately judge quantitative values from certain types of visual representations. As illustrated in Fig. 2.1, the most accurate visual encodings are spatial position (e.g., scatter plots), followed by length (e.g., bar graphs) and then angle comparisons (e.g., pie charts). This experiment has been reproduced and verified (e.g., [66]).

Much of the fundamental research in the psychology and vision research areas as they apply to visualization is nicely summarized by Ware [183, 184]. Ware explains that how humans process visual information have important implications for the perception and cognition of visualizations. For example, due to the basic anatomy of the human eye and the sensitivity of rods and cones, humans are more sensitive to brightness variations than color variations. Also, within the optical color spectrum, humans are more sensitive to red as compared to other colors. Regarding textures, visualization designers need to be aware of the possibility of interference patterns and, for example, need glyphs to have a very strong pop-out (i.e., a unique target among unalike distractors) visual encoding to overcome interference. Additionally, spatial proximity and connectedness have strong effects on the gestalt of a visualization.

These low-level concepts and their implications have both qualitatively and quantitatively been applied to design guidelines and principles in visualizations. For example, Edward Tufte highlights canonical visualization design principles in his books [173–175]. Some of these design guidelines include maximization of data density, maximization of the *data-to-ink ratio*, effective use of small-multiples to compare data or show changes over time, careful use of color, and strive to maintain *graphical integrity* (i.e., do not lie with the data).

In order to understand higher level cognition and interpretation of a visualization, more specific evaluations and case studies are needed. For example how effective is a visualization for accomplishing a task? Or for interpreting a particular type of data, how is a visualization understood by the user and how much cognitive workload and energy is required? To address these types of questions, evaluation methodologies as discussed in the next section are utilized.

2.2 Evaluation Methodology & Visualization Design

In order to evaluate the utility and effectiveness of visualizations, as well as measure fundamental perceptual and cognitive properties, evaluation studies are necessary. There has been extensive motivation and development of evaluation methods for use in visualization (e.g., [34, 75, 91, 99, 133, 170]). Using the terminology of [34, 116], the different anthropology and social science derived evaluation methods applicable to visualization and HCI include:

Field Study: Experimenter unobtrusively observes real situation in real environment; records observations.

Field Experiment: More intrusive version of a field study in which quantitative measurements are recorded and/or explicitly instructions or questions given to focus the observations.

Laboratory Experiment: Carefully designed experiment in artificial setting with carefully prescribed procedures. The behavior and observations can more easily be measured and quantified.

Experimental Simulation: An artificial environment is used to reproduce a field study or experiment in which the actual environment is either not practical or dangerous.

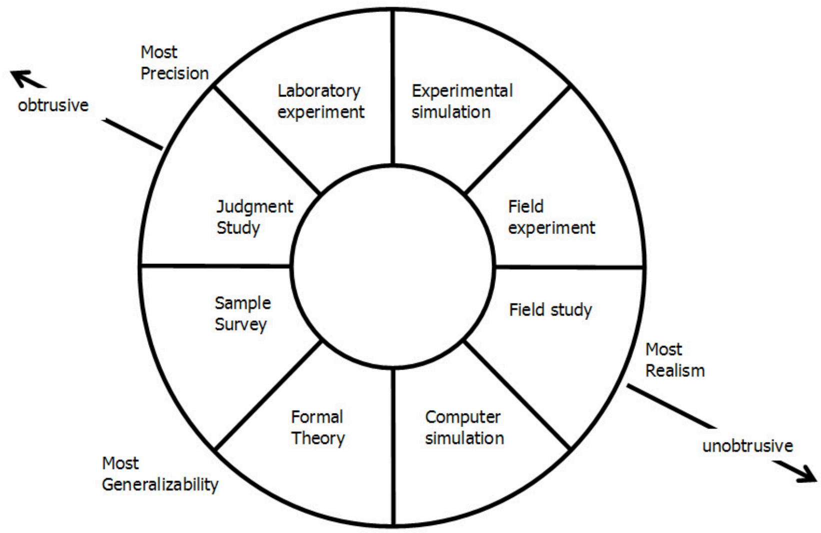

Figure 2.2: Illustrative diagram of visualization methodologies in relationship to their advantages and disadvantages. (Figure from [34], adapted from [116].)

Judgment Study: An artificial or neutral environment is used in order to measure a person's response to stimuli in a specific situation.

Sample Survey: A survey or questionnaire is presented to a given population in order to gain information from specific questions.

Formal Theory: Instead of conducting a new experiment or gathering new data, exisiting data and results are analyzed in order to extrapolate trends or new theories.

Computer Simulation: A computer simulation or model is used predict or simulate a situation commonly unknown or impractical for experimental situations. (Note: this methodology, unlike "Experimental Simulation", does not involve humans in the evaluation.)

Each of these methodologies have advantages and disadvantages in terms of generalizability of the results, domain task and environment ecological validity, and measurement precision, so must be carefully chosen for the given situation and evaluation needs. The trade-off's between generalizability versus specificity, as well as precision, are illustrated in Fig. 2.2. These methods can also be either *quantitative* (i.e., quantitative numerical data gathered as part of evaluation) or *qualitative* (i.e., textual or other non-quantitative data recorded during evaluation).

This thesis uses both quantitative and qualitative evaluation methodologies. In Chapters 3 and 4, we utilize the following methods: field study, field experiment, laboratory experiment, and sample survey. These methods were chosen in order to maintain ecological validity to the environment, users, and tasks evaluated with the domain specific problems as well as gain quantitative measurements to evaluate perceptual and cognitive properties. In Chapters 5 and 6, we utilize the following methods: laboratory experiment, judgment study, and sample survey. These methods were chosen as the motivating questions do not require a specific environment and primary goals are to gain precise quantitative measurements and generalizable results.

It should be noted that evaluations are instrumental not just in the study of perception and cognition of visualizations, and for visualization effectiveness, but are also a key part of the visualization design process. The place of evaluations in the visualization design process are nicely described within the context of *design studies* in visualization [120, 124, 157]. Evaluation of a visualization is part of the *validation* of a visualization: validation that it is usable and meets the task requirements of the users. In Fig. 2.3, the nine-stages of a design study as defined by [157] are illustrated. Evaluation methods are utilized primarily in the

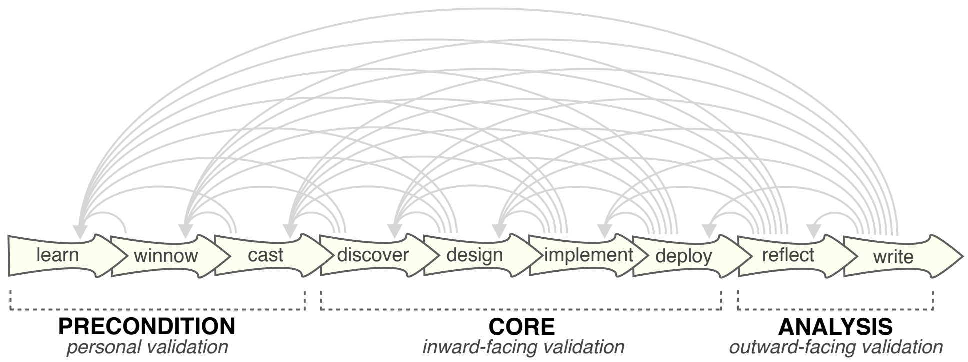

Figure 2.3: The nine-stages of a design study in visualization. (Figure from [157].)

inward-facing validation steps. In this thesis both Chapters 3 and 4 are examples of domain-problem driven design studies to develop a new visualization technique with evaluation methodologies employed to evaluate the tool as well as evaluate the perceptual and cognitive properties of the visual encodings.

2.3 Visualization in the Physical Sciences

Data visualization in the sciences and engineering is required for the effective analysis and communication of data. The sciences have a long history in data visualization and have been key to the field's development [49]. As the motivating visualization challenges in both Chapters 3 and 4 are primarily rooted in the physical sciences, and visualizations from scientific journal publications are explicitly discussed in Chapters 5 and 6, we briefly review the associated high-level key themes and visualization research agendas in this area. We review the more focused domain-specific related works within each of the chapters.

Within science domain journals, there have been a number of survey papers on visualization from, e.g., astrophysics [63, 84], geoscience [43], and chemistry [177]. There are also a number of general reviews of scientific visualizations (e.g., [58]) as well as reviews at the intersection of visualization and areas including medical imaging (e.g., [20, 143]), fluid flow visualization (e.g., [100, 117]), and vector fields (e.g., [130]). Looking at the physical sciences from the perspective of the visualization community, Lipşa et al. including M. Borkin [107] examined the past 10 years of publishing in computer science journals to identify trends and advances in visualization of the physical sciences as well as areas open for collaboration. The literature review utilizes a paper category classification scheme based on the major visualization challenges described in [81]. The papers surveyed fall into the following categories:

Multi-field visualization: The ability to effectively visualize multiple fields simultaneously so that it facilitates the analysis of the relations and correlations between those fields.

Feature detection: Locating features of interest in vast amounts of data, representing them and tracking the evolution of features in time and/or space.

Graphics hardware: Novel methods to harness available graphics hardware (GPUs) to address large scale and real time rendering.

Modeling and Simulation: The use of scientific knowledge in specific fields to model, simulate, and visualize data.

Scalable visualization: Research methods to address challenges created by large data: I/O, processing, and visualization.

Error/uncertainty visualization: Integration of error visualization into the main visualization of data.

Time-dependent visualization: Methods for visualizing time dependence and changes over time for given data.

Global/local visualization (details within context): The techniques in this category aim to integrate a visualization of the whole data required for navigation and a global understanding of the phenomenon described with selection and detailed visualization of sub-sets of interest.

Comparative visualization: Comparative visualization refers to the process of understanding how data from different sources are similar or different.

Interestingly, there are *no* physical science visualization papers published in the past 10 years addressing the following challenges:

Perceptual issues: The study and application of knowledge of the human visual system to the design of visualization techniques.

Quantify effectiveness: Evaluation and comparison of new to previous techniques, and quantification of their effectiveness through user studies.

Human-computer interaction: Development and implementation of effective interaction with visualizations.

Integrated scientific and information visualization: Techniques to handle high-dimensional information and the compositing of both information (i.e., non-spatial) with scientific (i.e., spatial) visualization methods.

Visual abstractions: The creation of visual abstractions, in the flavor of information visualization, to more effectively understand and analyze scientific data.

Theory of visualization: Theorization of the practices of visualization.

As stated in [107], "Some of these challenges (perceptual issues, quantify effectiveness, human-computer interaction, visual abstractions and theory of visualization) are mostly addressed by papers not associated with a physical science." Later in the paper it states, "There is also potential for research in [these] other visualization problem categories." In this thesis we do indeed address the visualization challenges at the intersection of all these under-explored areas, primarily in Chapters 3 and 4, with motivating examples, implications, and relation to the physical sciences as well as other areas of biomedicine and engineering.

In the next two chapters we will address two case studies, one from biomedicine and one from engineering with broad applicability to scientific applications, and address both the novel visualization solutions created as well as the perceptual and cognitive evaluations results with broad generalizable results.

3

Evaluation of Artery Visualizations for Heart Disease Diagnosis

This chapter presents the first of two domain-specific case study evaluations in this thesis. As discussed in Sec. 2.2, fundamental perceptual and cognitive theories can be tested for accuracy and completeness through such case studies. New cognitive theory can also be derived for specific types of tasks. In this chapter we will discuss a case study in biomedicine, the new visualization tool and technique

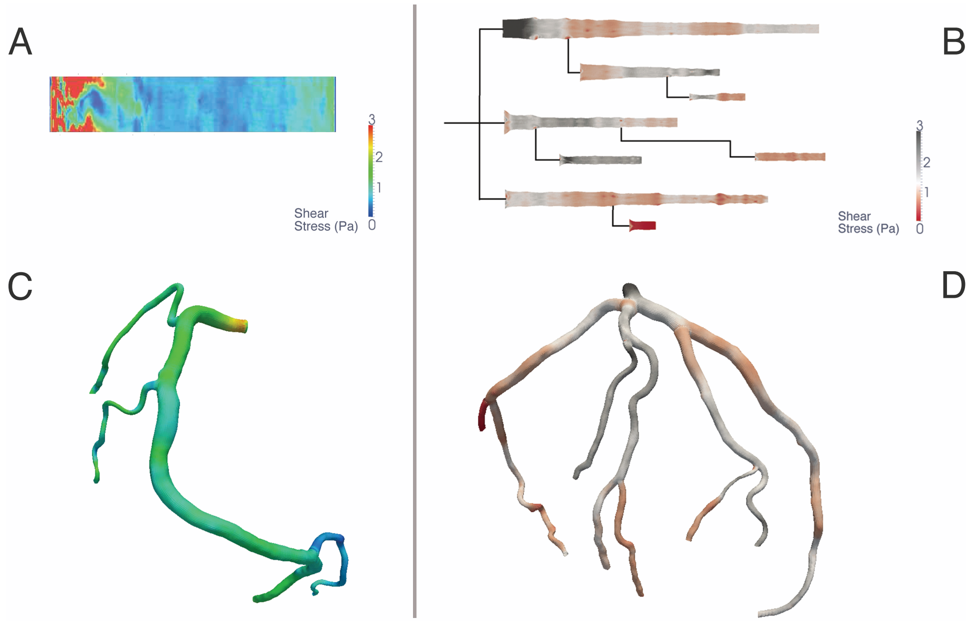

Figure 3.1: *Left:* Traditional 2D projection (A) of a single artery, and 3D representation (C) of a right coronary artery tree with a rainbow color map. *Right:* 2D tree diagram representation (B) and equivalent 3D representation (D) of a left coronary artery tree with a diverging color map.

developed as part of the project, and the results of a user study evaluation.

3.1 Motivation

In the United States, the leading cause of death is heart disease resulting in over 600,000 deaths per year [109]. Early prevention and treatment is vital for saving lives, and visualization plays an essential role for patient diagnosis in cardiovascular imaging. A new non-invasive diagnostic technique under development uses Computed Tomography Angiography (CTA) data from patients combined with blood flow simulations to calculate hemodynamic risk factors, in particular Endothelial Shear Stress (ESS), in coronary arteries [153]. Visualization methods of this data are of great value for this emerging research and have the potential to lead to faster, more accurate heart disease diagnoses.

Current visualizations techniques, as shown in Fig. 3.1 (left), use either a 2D cylindrical projection of a single artery or a 3D representations of the coronary artery tree. In both cases, ESS is mapped to the surface using a color encoding, typically with a rainbow (or "spectrum") color map. Both representations have their advantages and disadvantages: 2D allows one to see all the data at once, but anatomical information is lost both in the shape of a vessel and in how each individual vessel connects to other branches. 3D preserves the anatomical structure, but introduces issues of occlusion and requires human interaction to rotate the model in order to see all the data. The fundamental visualization issue is how to display a scalar quantity – ESS – that is parameterized on the surface of a three-dimensional spatial structure – coronary artery trees – using effective visual encodings.

In collaboration with doctors and researchers in cardiovascular imaging and applied physics, we set out to investigate the effectiveness of different visualization strategies for this problem. We first conducted an informal qualitative user study with our domain experts to develop a task taxonomy and determine their current visualization practices and needs. During this process we developed HemoVis, a novel 2D tree diagram representation that presents all the data at once while still encoding pertinent anatomical information such as vessel circumference and branch structure (Fig. 3.1, B). Despite positive feedback about HemoVis, we encountered resistance from users to work only with the 2D representation and to use a diverging color map that emphasizes features in the data more effectively. To investigate these issues we conducted a formal quantitative user study with medical domain experts to evaluate the effectiveness of 2D versus 3D representations, and the effect color has on task completion performance.

The first contribution of this chapter is a task taxonomy for hemodynamics that is based on a qualitative user study with medical experts. Guided by this task analysis, our second contribution is the design of HemoVis, an interactive visualization application for heart disease diagnosis that uses a novel 2D projection and layout for artery trees. The third and main contribution of the chapter is a quantitative user study with domain experts that shows statistically significant results demonstrating that our 2D representations are more accurate and efficient than 3D visualizations. In addition, the performance in 3D drops with an increase in the complexity of what is being examined while the 2D representation is insensitive to it. Our study also shows that a perceptually appropriate color map leads to fewer diagnostic mistakes than a rainbow color map both in 2D and in 3D, and that task completion times were nearly twice as slow with the rainbow color map in 3D. To the best of our knowledge, this is the first quantitative evaluation of the effect of spatial and color encodings for a domain task with medical professionals and real patient data.

3.2 Related Work

Vessel visualization: Vessels or branched systems are primarily visualized with 3D representations [29, 57, 125, 135] with scalar quantities such as ESS mapped onto the surface using color encodings [48]. In 2D, a common technique is to generate a Curved Planar Reformation (CPR) visualization of the vascular structure, where an image is generated by taking slices along a generated centerline and displaying the intensity units [83, 164]. A drawback of CPR is that it provides limited information about the artery wall, thus not giving a clear indication of how wide or narrow the vessel is at any particular point nor convey ESS data for the entire artery. There are also other 2D projection techniques all with the goal of flattening the structure into a single view [150, 185, 186]. 3D representations have the problem of occlusion that does not allow a doctor or researcher to see all the data simultaneously. Techniques have been developed to improve 3D representations using visual cues such as shadows and transparency to improve spatial acuity in 3D [97, 142] or to include 2D data [166]. In our 2D representation, all the necessary data including ESS is displayed for a medical practitioner to assess the severity of disease within an anatomical frame of reference.

2D vs. 3D visualizations: Formal evaluations are a valuable measure to determine the effectiveness of visual representations and data encodings [91, 133]. A variety of case studies and formal user studies have demonstrated that 2D data encodings and representations are generally more effective than 3D for tasks involving spatial memory, spatial identification, and precision [38, 172]. Excellent examples exist in the realms of vector field [47, 98, 114] and geospatial visualization [5, 19, 87, 93, 131]. Although 2D is typically more effective, there are strategies to improve 3D performance, such as using occlusion and perspective [160, 182] or stereographic displays [171]. Our work investigates the effectiveness of 2D versus 3D artery representations in the context of heart disease diagnosis by medical experts.

Color map evaluations: Choosing the appropriate color map is essential for the effective display and analysis of quantitative data. Based on fundamental human perceptual principles and the type of data being displayed (sequential, diverging, or categorical), there are formal and systematic ways to make an appropriate color choice based on the task at hand [64, 122, 141, 148, 168]. Based on laboratory user studies, specific guidelines are available for the effective design of color maps [140, 149, 181]. A particular color map of interest is the rainbow map which, despite being a favorite color map across the sciences [26], is poorly suited for most data tasks and can prove misleading since it is not perceptually ordered and isoluminant [26, 106, 147, 148, 181]. Quantitative studies confirm these facts [86, 146] and propose better ways to design color maps and discern when isoluminant maps are suitable. Despite this general body of knowledge, there has been little study on the effects of the color map within a real-life domain application. We present a quantitative evaluation of the rainbow map's effects on task accuracy and efficiency by domain experts within a real domain application.

3.3 Scientific Background

Atherosclerosis, the disease focused on in this research, occurs when plaque forms in the arterial wall, causing possible obstruction of blood flow as well as changes in the outer dimension of the artery. Sites of plaque deposit (*atherosclerotic lesions*) form where the endothelial cells that line the arterial wall exhibit increased inflammation and permeability to lipid molecules such as LDL (i.e., bad cholesterol). Over time these plaques can either become a low-risk type that is quite large and causes a narrowing of the artery (*stenosis*), or a high-risk type that can rupture, potentially causing a heart attack. These high-risk deposits are not detectable with conventional imaging. In the United States, approximately 300,000 deaths per year occur from coronary artery disease and $\sim$60% are caused by rupture of these high-risk plaques [46]. However, recent research has shown that areas of low endothelial shear stress (ESS), i.e., the frictional force of blood on the artery wall, stimulate the development of these high-risk lesions [36, 163], and that these lesions primarily appear where there is disturbed flow, e.g., at artery bifurcations, bends, and regions of increased diameter. Thus ESS is a powerful indicator of plaque formation and disease progression.

However, it is impossible to directly measure ESS *in vivo* for an entire arterial tree. As a consequence, one needs to rely on blood flow simulations to calculate a patient's ESS based on their artery geometries. Combining this blood flow simulation with a patient's 3D reconstruction of their coronary arteries allows doctors to detect areas of low ESS, identify plaque sites non-invasively, and take preventative measures before a heart attack occurs [153]. In order to have ecological validity and develop visualizations specifically targeted at the most important diagnostic tasks, we conducted a formative qualitative user study to determine a user's tasks, what data needs to be visualized to perform these tasks, and what are the best ways to visualize the required data.

3.4 Formative Qualitative User Study

3.4.1 Overview & Logistics

The first goal of this study was to characterize the medical and research problems being addressed by the participants and their specific domain tasks related to atherosclerosis. This was achieved by conducting a series of semi-structured interviews with 10 medical doctors and researchers representing the potential future users of such data. All the participants were interviewed at and affiliated with Brigham and Women's Hospital (Boston, MA). In an attempt to cover as broad an audience as possible, study participants ranged in age, gender, experience level, education background, job seniority, clinical versus research focus, and department (radiology versus cardiology).

Each participant was interviewed and asked the same set of questions to gather sufficient information on their background and experience, research interests, knowledge of hemodynamics, and their current workflow goals and tasks. Each participant was then shown a series of images covering 2D and 3D representations of ESS. The desired outcome was, based on the resulting feedback, to answer the following questions: What data should be shown to accomplish tasks of clinical importance? What are the optimal 2D representations? What are the optimal 3D representations? Should the data be encoded in 2D or 3D? And what color schemes are best to aid the individual in task completion?

3.4.2 Task Taxonomy

The participants' jobs fall into two broad categories: clinical diagnostics and fundamental research. The former represents those individuals who work on making clinical diagnoses for specific patients, and the latter represents individuals who work on investigating the fundamental causes of heart disease. Some individuals fall into both categories (e.g., doctors who split their time between medical practice and research). Table 3.1 represents all the domain tasks cited by participants and category of individuals that cited the task. Both categories of individuals need to accomplish the same basic set of tasks essential to the immediate diagnosis of patients based on factors that have a proven link to atherosclerosis (i.e., low ESS, artery geometry). However, those with a research focus care about further exploring the data to investigate variables or aspects not necessarily with a proven link to disease.

Each of these domain tasks can be abstracted to fundamental analytic tasks

Table 3.1: Domain tasks, broken down by clinical versus research focus, of participants based on a formative qualitative user study.

ID	Task	Clinical	Research
1	Identify stenosis or blockage	X	X
2	Identify regions of low ESS	X	X
3	View all ESS data for heterogeneous patterns	X	X
4	Study blood flow (velocity) patterns		X
5	Identify regions of blood recirculation		X
6	Investigate other physical variables of blood flow		X
7	Follow patient's disease progression	X	X

presented by Amar et al. [2]. Tasks 1 and 2 correspond to *finding extrema*. The clinical diagnostics study participants expounded on the need for these tasks to be accomplished in a quick and efficient manner with minimal cognitive effort for rapid diagnosis of the patient. These tasks are usually done inside or just outside a procedure room to expedite taking preventative measures (e.g., insertion of a stent). In a research setting the time pressure was lower. Task 3 corresponds to *clustering* of the data. In a clinical setting, this task is not as essential or of the same importance as 1 or 2, but would be necessary for diagnosis in complicated cases. Task 4-6 are about *finding anomalies* in the data. In these cases time is not a factor, and being able to analyze the data carefully is the primary goal. Finally, task 7 is a *correlation* task, requiring comparison of multiple data sets, if available for a patient, in order to follow the progression of the disease.

3.4.3 2D Representations

Projections: The first portion of the qualitative user study was focused on 2D projections of the arteries. This is relevant to Tasks 1-3 (see Table 3.1). All but

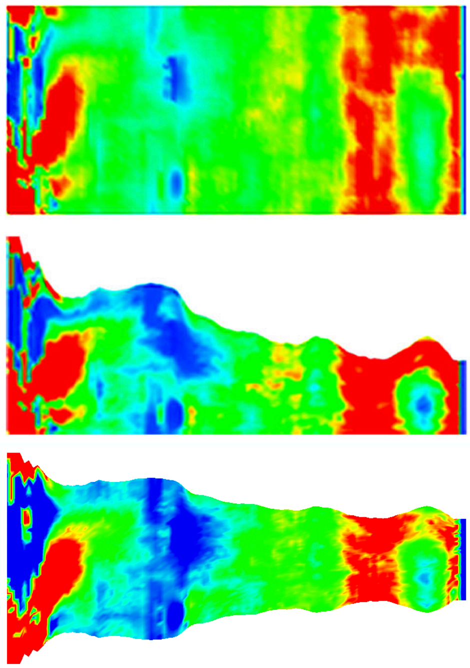

Figure 3.2: The three different cylindrical projection techniques presented to users during the formative qualitative user study. *Top*: Traditional cylindrical projection. *Middle*: Projection with circumference mapped to image height above the x-axis. *Bottom*: Projection with circumference mapped to the width of image symmetric on centerline (preferred projection by users).

one of the participants had been exposed to reading ESS values from a cylindrical projection representation of an artery, and all were familiar with ESS in general. Initially two different types of projections were shown to participants (Fig. 3.2 top and middle). The top image is a traditional cylindrical projection in which the length of the image is based on the length of the artery, and the width is arbitrarily chosen. The middle image represents a variation on this cylindrical projection in which the width equals the circumference of the artery at the cross section.

During the user study, the first two participants in the post-evaluation interview commented on the middle cylindrical projection. They both said that it was slightly confusing to them and that, when viewing the image, they were mentally reflecting the image over the x-axis to make it symmetric. Upon presenting them with the pseudo-cylindrical projection based on their feedback depicted in Fig. 3.2 (bottom), they both commented that this was a superior mapping. This projection maps the circumference of the artery at the cross section to the height, and centers it along a centerline. For the remainder of the study we presented all three projections to the participants.

All the participants preferred the pseudo-cylindrical projection map (Fig. 3.2 bottom). When asked whether this projection is confusing, they replied "no". Although not exactly reproducing the real geometry, the representation is able to encode enough geometric information to show the user the relationship between geometry and ESS as well as where stenoses occur. This representation is also intuitive for clinicians since the 2D mapping mimics what one would see if an artery is cut-open ("butterflied") for *ex vivo* studies.

Tree Diagram Layout: Continuing with 2D representations, for tasks 1-3, all of the study participants were presented with a visualization of the artery tree in with each node was representative of an artery and branches indicative of bifrucations (Fig. 3.3). When presented with the hierarchical tree layout in Fig. 3.3 (top), no one had ever seen anything like it before and all except one individual responded that they greatly preferred this data representation and found it useful. Unlike CPR visualizations that only include two locations of the artery wall [83, 164], this tree diagram not only shows the artery structure but conveys the length and width of each vessel. The most common feedback on the tree diagram was the great benefit of being able to see all the data at once without the usual occlusion challenges found in 3D representations, and being able to display and

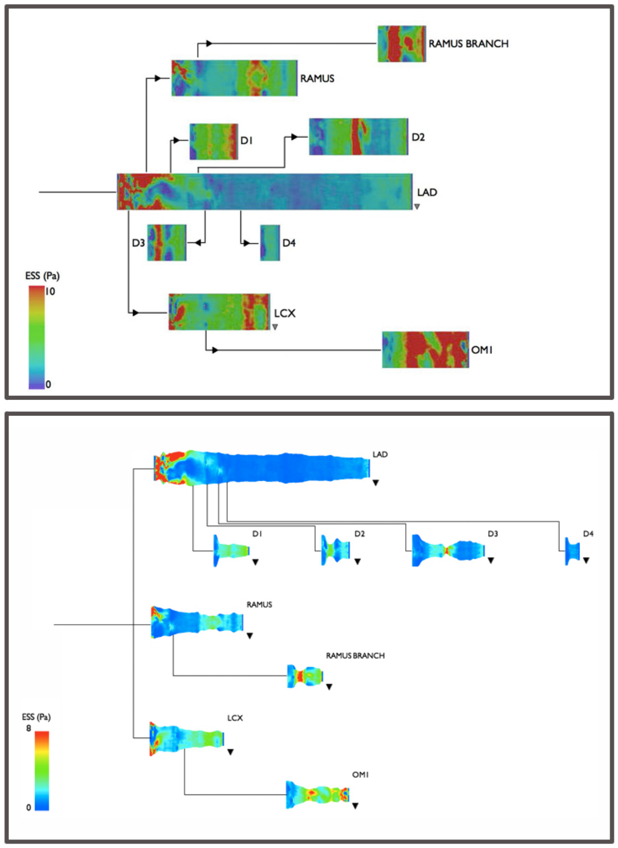

Figure 3.3: Original and final tree layout schemes. *Top*: Initial design of tree layout. *Bottom*: Final tree layout based on feedback from users to make it more anatomically representative.

compare multiple data sets at the same time (task 7). The two most commonly cited scenarios where this would be useful are for viewing a patient's data with the blood flow simulation adjusted to simulate a patient at rest and at heightened physical exertion (i.e., low and high flow rates), and for viewing multiple image acquisitions of a patient over time in order to view and compare the simulation output for each. This latter scenario would aid the doctor in observing the progression of disease over time.

As shown in the initial sketch (Fig. 3.3 top), the study subjects were presented with some possible interactions with the tree data including the ability to open and close branches, move branches around the screen, and interactively crop the color scale. All of the participants said that closing the branches is not useful since they always want to see the other data sets and keep the data in context. Everyone also stated that being able to move or drag branches around the screen is not a good idea since they want to maintain the same static anatomical structure when viewing the data for context.

All of the user study participants, except the research staff participants without advanced degrees, noted that the tree was not accurately conveying the anatomical information. When the initial design (Fig. 3.3 top) was created, the branches were arranged to fit the space available on the screen. Independent of each other, the participants were asked to describe or draw what adjustments they would make to create an anatomically "correct" version of the tree. All of these individuals made the same recommendations. For the vertical arrangement, the branches should be ordered from major to minor arteries with the placement of each descending branch point below the "parent" vessel (and if descending branch "superior", i.e., branch points anatomically go upwards, place it above the parent vessel). If there are two very close major bifurcations, then present it in the tree diagram as a trifurcation. Finally, always have the tree diagram lines angled and drawn in the same direction as the blood flow (does not make sense to have blood flow "upstream") which also eliminates the need for directional arrows.

The final diagram is shown in Fig. 3.3 (bottom). It should also be noted that the sample data set we used for the study has a large complex branched structure and that the ESS data is high resolution compared to a typical data set. This was chosen on purpose to represent the most complicated large case available with conventional scan technology. Finally, to evaluate whether a portrait or landscape presentation was more useful, the participants where shown a tree diagram representation in each orientation. All the participants preferred the landscape orientation depicted in Fig. 3.3 (bottom) since it was intuitive to “read” the visualization and flow direction from left to right, and comparisons were easier for multiple data sets when stacked vertically.

3.4.4 3D Representations

The next portion of the formative qualitative user study was focused on the 3D display of ESS data, and which of 2D or 3D display techniques may be better for the completion of the domain tasks. All of the participants had previously seen 3D representations of ESS as a 3D surface with ESS mapped in color onto the exterior (see Fig. 3.1, C). Mapping ESS in 3D is useful since it maintains the true geometry and anatomy a doctor is accustomed to seeing, and makes it easier to translate the knowledge back to a surgical setting. A common practice, cited by the study participants and visible in the literature, is to choose ⊠standard⊠ viewing angles that a doctor or surgeon would be interested in. However, to view all the data effectively at the same time requires interactivity or animated rotation. 2D projections of ESS, such as those discussed in Secion 3.4.3, have the advantage of being able to show all the data without occlusion, to display all the data in a static state, and to easily compare multiple data sets.

3.4.5 Color

A portion of the qualitative study was devoted to determing the best color schemes to encode scalar ESS data in either 2D or 3D. Except for this portion of the study, all images presented during the study were created using a rainbow color scheme. Rainbow is the standard color map used in the medical literature, and we did not want the participants to be distracted by unusual colors when asking about spatial data encoding techniques.

A total of eight carefully designed simple color schemes were presented to the users (see Fig. 3.4). The priority was designing a scheme that brought out the data structure to accomplish tasks 1-3, particularly the data which is not as visible when viewed with the standard rainbow map [105]. Four of the color schemes are based on diverging color tables from Cynthia Brewer's ColorBrewer [28]. The other two diverging color schemes were designed based on the concept of luminance being easy to read with it encoding the scalar ESS value and with chroma only highlighting the highest or lowest ESS values. Also included was a simple luminance scheme with no color. The color schemes presented to the study participants are shown in Fig. 3.4: a rainbow map [A], a desaturated rainbow map with yellow at the midpoint (rather than the standard green) [B], a diverging map from blue to red with white at the divergence point [C], a diverging map from blue to red with cream at the divergence point [D], a sequential greyscale map [E], a diverging map from red to black with white at the divergence point [F], a diverging map from blue to black with white at the divergence point [G], and a diverging map from purple to orange with white at the divergence point [H].

Every study participant except one said that they liked the rainbow scheme the

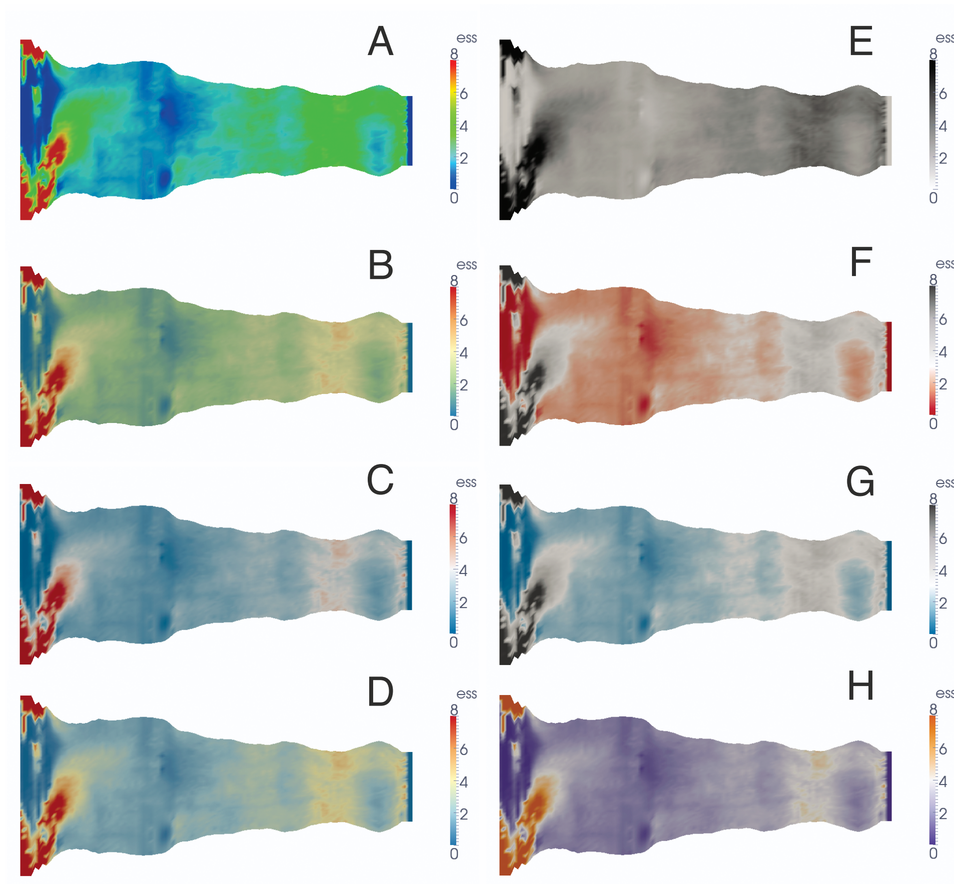

Figure 3.4: Color schemes presented during the qualitative user study. The rainbow scheme (A) was preferred by most since it is what they are accustomed to viewing. The next most popular scheme was the red-black diverging scale (F). The grayscale image (E) was unanimously disliked since participants assume black-and-white images to be raw radiological data, while color indicates that the data has been processed or simulated.

best. The reasons for this are that it is what they are "used to seeing", the colors are more saturated than those in the other scales making it "easier to see", it is the "most aesthetically pleasing" of the choices, and that it is the "easiest to directly determine the numerical value" of the ESS based on matching the color to the scale. However, the participants were astute and saw, after viewing all the color maps, that red is the most eye-catching as it has a "pop-out" effect [13]. Multiple study participants acknowledged that their primary task is identifying regions of low ESS so it would make sense to have the areas of concern be red. Thus they suggested a more useful version of the rainbow scale would be to invert it, so that red indicates low ESS and blue high ESS.

Examining the non-rainbow schemes, the simple grayscale shows great detail and subtleties in the ESS data, especially compared to the rainbow scale. However, most of the participants objected to this scheme citing that when they see a black-and-white image they assume the data is raw radiological imagery (e.g., x-ray, MRI, etc.) and not simulated data. When they see a false-color scale, they immediately make the assumption that it is processed and/or simulated data and not raw imagery. Thus it is important to include some form of color when mapping data like ESS to avoid confusion.

Of the non-rainbow schemes, the one study participants liked the best was the red-to-black diverging scale. They felt it did the best job of grabbing their attention to the highlighted areas of extreme ESS and showing the data structure. Although some of these users had pointed out that a pure luminance scale with no color usually indicates raw imaging data, none of them complained or mentioned this when picking the red-to-black scale. Finally, one user liked the diverging red-to-blue schemes the best. Also, a number of participants gave positive comments during this color design portion of the study, acknowledging that they could see more structure in the data when using a non-rainbow scheme, and that they would consider using a diverging color scheme instead in their own data analysis.

In summary, the key take-aways and lessons learned during the formative qualitative user study are to keep the data representation as anatomically correct as possible (i.e., choice of 2D projection and tree diagram layout); that a 2D data display is more effective for data analysis and for comparing multiple data sets; that the best color choice is a diverging color scheme utilizing red to highlight the regions of greatest interest; and that a pure black-and-white color scheme should be avoided since users associate it with raw radiological data. As will be discussed in the following section, we applied all these principles in the development and design of HemoVis.

3.5 HemoVis

Using an iterative task-driven design based on our formative qualitative user study, we developed a 2D interactive visualization called HemoVis[1] (Figs. 3.5 & 3.6). The design is based on the qualitative evaluation with additional feedback from select users and the task taxonomy with a focus on the tasks that are most relevant for both clinical and research settings (1-3 & 7 from Table 3.1) as described in Sec. 3.4.2.

HemoVis has two viewing modes: *tree* (Fig. 3.5) and *individual* (Fig. 3.6). In tree mode, a tree diagram of the arterial system is presented in which each node is representative of an artery and each line segment representing a bifurcation.

[1]Available online at http://www.seas.harvard.edu/∼borkin/HemoVis

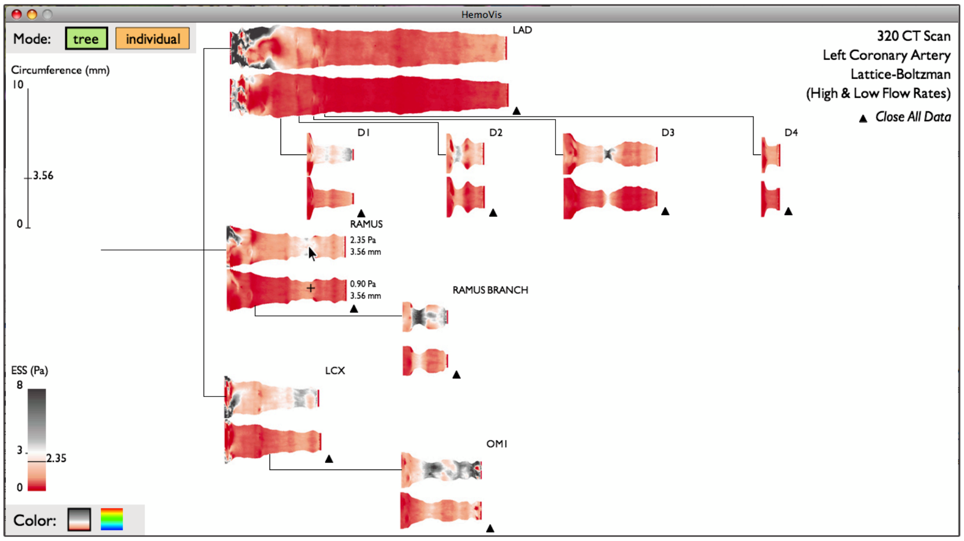

Figure 3.5: HemoVis in the "tree" mode displaying a patient's left coronary artery tree with color mapped to ESS.

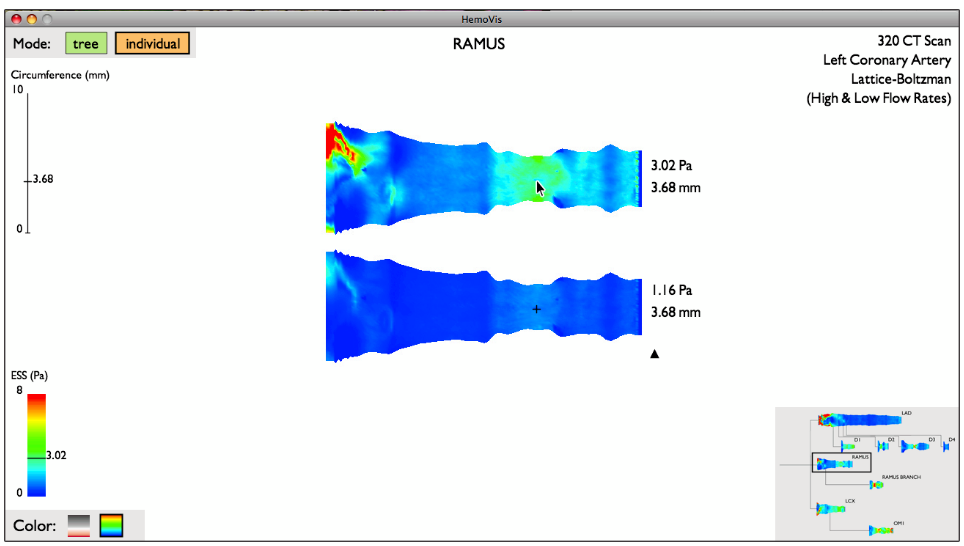

Figure 3.6: HemoVis in the "individual" mode displaying a single artery with small tree diagram for navigation in the lower right.

Each artery is displayed using the 2D pseudo-cylindrical projection discussed in Sec. 3.4.3, is labeled with its anatomical name, and has its ESS values encoded with color. The color and size scales are displayed to the left of the tree diagram. The upper right of the screen displays the relevant metadata for the particular data set. The interaction techniques implemented are based on expert feedback from the qualitative study and follow-up discussions. The user is able to simultaneously view additional simulation data sets for the same patient by clicking the small triangles. A user can mouse-over the images to display exact quantitative ESS and circumference measurements. If the second data set is open for a particular artery, then a cross-hair cursor will appear on the image opposite from the mouse cursor indicating the equivalent position. One can also change the desired color mapping by selecting one of the colored boxes in the lower left corner. To switch modes, there are two mode buttons in the upper left. In the alternate individual mode, only one artery is displayed at a time allowing the user to take care at studying particular arteries in high resolution. In this mode there is also a small non-interactive version of the entire coronary tree in the lower right corner to help keep the displayed artery in context as well as be used to navigate the branches by clicking them. HemoVis is implemented in Processing [2].

The data for both our qualitative and quantitative (Sec. 3.6) user studies comes from the Multiscale Hemodynamics Project[3]. The patients' coronary geometries are obtained from CTA data acquired with a 320 detector row Toshiba AquilionONE scanner. The data is 4D from a series of cardiac cycles which is then registered into a single volume. The data is then semi-automatically segmented

[2]http://www.processing.org

[3]http://hemo.seas.harvard.edu

using Vitrea (Vital Images Inc). The end result is a series of 3D surfaces of the heart and coronary arteries. These geometries are then loaded into MUPHY, a multi-physics and multi-scale code combining Molecular Dynamics (MD) with a Lattice Boltzmann (LB) method, to model the blood flow through the static geometries [118]. The simulation was run using a parallel implementation on Harvard's IBM BlueGene/L. The result is 3D data of the simulated blood flow and associated properties including ESS.

However, despite positive feedback from users and the domain expert driven iterative design, many potential users were reluctant to try the prototype because they were not convinced it was really better than a 3D representation. Also, despite expert acknowledgment during the formative qualitative study's section on color choice that some of the non-rainbow schemes did an excellent job of displaying features in their data and presenting users with background literature on the rainbow color map (see Sec. 3.2), the users were reluctant to choose an alternative color map to rainbow. Additionally we wanted to quantitatively investigate, with this real world example, what effect data representation and color encoding has on task performance. As a result, we decided to conduct a formal quantitative user study to evaluate if a 2D representation is more effective than 3D and if color can effect how one perceives ESS features.

3.6 Quantitative User Study

We conducted a formal quantitative user study to determine whether a 2D or 3D data representation of ESS was more effective and efficient for diagnosing a patient's coronary artery disease. We also wanted to see if there were quantitatively measurable performance effects based on the color scheme utilized, specifically the rainbow and the diverging red/black color maps. Having an effective visualization is important in making an accurate diagnosis, but having an efficient visualization is also important in order to allow a medical professional to take rapid preventative measures if needed as well as increasing overall hospital efficiency. To maintain high external validity of our results for this domain, we worked with medical professionals and real patient data.

3.6.1 Hypotheses

Our hypotheses entering the user study were:

H.1 Compared to a 3D representation, a 2D data representation will result in fewer diagnostic errors and faster performance.

H.2 A non-rainbow color map, specifically a diverging color map, will result in fewer diagnostic errors and faster performance than a rainbow color map.

3.6.2 Participants and Apparatus

In order to have a large number of medically literate participants in the study, we chose to use medical students. These participants all had the basic medical expert training and knowledge of cardiovascular disease and anatomy and did, therefore, have the necessary expertise to fully understand and complete the tasks presented in the study. Participants had no prior bias towards any specific ESS visualization because the non-invasive diagnostic technique being presented here is not yet a part of standard clinical practice.

Twenty-one Harvard Medical students participated in the study. This included 12 women and 9 men, with a spread of 1^{st} through 4^{th} year students. All participants reported having normal color vision and were additionally checked at the beginning of their session using a standard Ishihara pseudo-isochromatic plate series for detection of protan/deutan (i.e., red/green) and tritan (i.e., blue/yellow) color vision deficiencies. Each participant was monetarily compensated for their time at the end of their session.

All study sessions were conducted in the same room with identical lighting, and on the same MacBook Pro 15" laptop. Participants where offered the choice of a wireless mouse or trackpad based on which they felt more comfortable using; every participant chose the wireless mouse. The 2D representations were displayed using Preview, and the 3D representations were displayed using Paraview. Audio, video screen capture, and mouse clicks and movements were all recorded with Screenflick.

3.6.3 Tasks and Procedure

The study session for each participant started with the color vision test, followed by a basic survey to obtain demographic information and to assess their knowledge of both heart disease and fluid dynamics. The participants were then provided with two pages of background information giving a brief overview of the project and the new non-invasive diagnostic tool being evaluated. Next the participants where given instructions for their task ("identify all low ESS regions") and shown a series of images (see Fig. 3.1 (B) for sample 2D and (D) for equivalent 3D representation) for them to perform the task on with a survey in-between each image to gauge their confidence levels. The session concluded with verbal questions and feedback.

During the main part of the experiment, each participant was shown on a LCD screen a series of 8 images with the first 2 serving as a training tasks. The images alternated between 2D and 3D representations to minimize difference in learning effects between the two representations, and the images included an evenly distributed mix of left and right coronary artery trees (since the left and right sides have slightly different anatomical structures and complexities as demonstrated in the 3D representations (C and D) of Fig. 3.1). The data used in the 3rd and 4th images shown to the participant were also used in the 7th and 8th images but using the alternate 2D/3D representation to allow for a larger number of measurements per participant.

The participant's task (which was explained in both printed and verbal instructions with annotated sample visualizations) was to identify all the low ESS regions in each image. In both 2D and 3D conditions, a person could indicate small regions by clicking on them with the left mouse button, while larger regions could be marked by encircling them with the mouse cursor. These actions left no visible mark on the image, but were recorded by our software for post-experimental analysis. In 3D, the participant could arbitrarily rotate the model. Based on results from our pilot run of this study, we did not enable zooming because it did not improve the diagnostic accuracy, but frequently caused participants to become disoriented, losing track of which parts of the image they had examined and which they had not.

After each image, the participants filled-out a questionnaire where, based on the task they had just completed, they were asked to respond on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) to four statements: "I found it easy to identify low shear stress regions", "I was able to perform the task efficiently", "I am confident I found all the low shear stress regions", and "I am confident all the places I marked are really low shear stress."

At the end of the session, each participant was verbally asked which visualization style (i.e., 2D or 3D) they preferred and why, and asked whether they had other comments, questions, or feedback. Each session lasted approximately 40 minutes.

3.6.4 Experimental Design & Analysis

The study was a mixed between- and within-subject design with the following factors and levels:

- dimensionality of representation (2D or 3D);
- color mapping (rainbow or diverging).

Dimensionality of representation was a within-subject factor and color mapping was a between-subject factor.

Our dependent measures were the fraction of low ESS regions identified, the number of false positives (i.e., non-low ESS regions identified as ESS), and the time to complete a diagnosis. Because the time to complete a diagnosis was impacted by the number of low ESS regions a participant identified in each image as well as the total number of low ESS regions present in each image, we additionally compared participants' performance in terms of the average amount of time taken to identify a low ESS region (i.e., total time spent on an image divided by the number of low ESS regions identified).

To generate the two accuracy-related measures, each participant's responses (regions they encircled with the mouse cursor or clicked on) were compared against answer keys generated by cardiovascular imaging specialists. After each trial, we also collected four subjective measures as described in the previous section.

Half of the participants started with a 2D representation and half with a 3D representation. Similarly, half of the participants started with a left coronary artery tree and half started with a right. The orderings of data sets (three hearts) were counterbalanced using Latin Square design. Genders were balanced between the two color mapping conditions and between tasks starting with 2D and 3D representations.

The time to complete a diagnosis followed a lognormal distribution. We log-transformed these data as is common practice and analyzed it with a t-test. For the remaining measures, we used non-parametric tests: the Wilcoxon signed rank test for within-subject comparisons, and the Mann-Whitney U test for between-subject comparisons. To guard against Type I errors, we applied the Holm's sequentially-rejective Bonferroni procedure [158] to the analyses of the subjective responses and to the additional analyses that did not correspond directly to our two stated hypotheses.

Because of the substantial qualitative differences between the 2D and 3D conditions, we analyzed the effects of color separately for each of these two conditions.

3.6.5 Results

Preliminaries: A contrast analysis of the fraction of low ESS regions identified across the 6 test tasks revealed no significant learning effects ($Z = -21.5$, $p = 0.47$)[4]. That is, the participants' ability to correctly identify low ESS regions did not change significantly throughout the experiment. We thus include results from all 6 tasks in our subsequent analyses.

[4]Z = z-score for the Wilcoxon signed rank test, p = p-value.

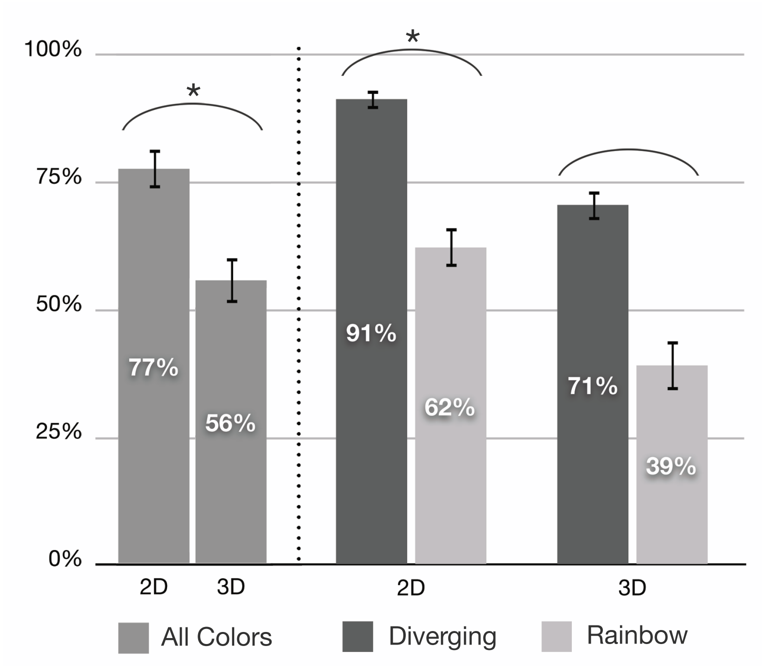

Figure 3.7: Average percent of low ESS regions identified broken down by 2D and 3D representation, and color. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. Participants were more accurate in 2D and when using the diverging color map.

Accuracy: We observed a main effect of the dimensionality of representation on the fraction of low ESS regions identified ($Z = -115.5, p < 0.001$): participants correctly identified 77% of low ESS regions in 2D images, but only 56% in 3D (see Fig. 3.7). In both 2D and 3D conditions, we also observed significant effects of color mapping on the fraction of low ESS regions identified. For 2D images, participants in the diverging condition found $\sim$47% more low ESS regions than the participants in the rainbow condition ($U = 1$, $p < 0.001$, $r = 0.83$)[5]. For 3D images, the diverging color map resulted in an $\sim$82% improvement over the rainbow color map ($U = 7$, $p = 0.001$, $r = 0.74$).

[5] U = Mann-Whitney U measure, p = p-value, r = Z / $\sqrt{N}$

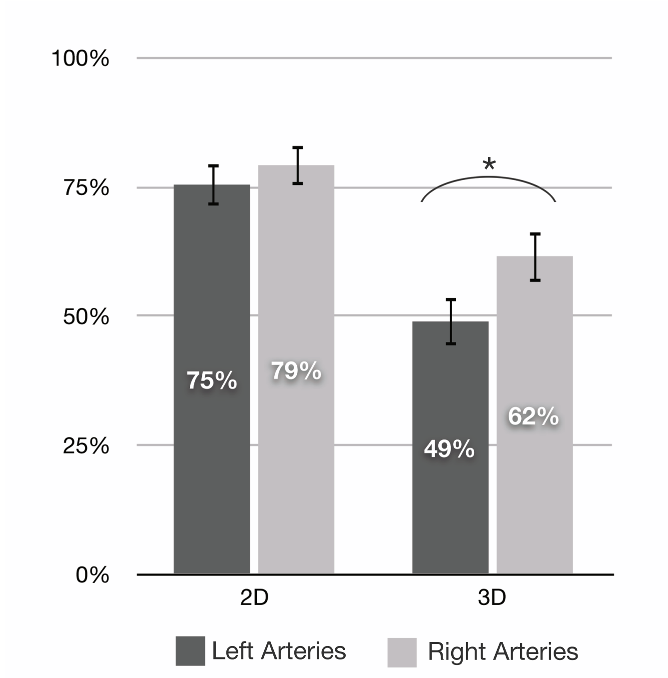

Figure 3.8: Average percent of low ESS regions identified broken down by 2D and 3D representation, and left and right artery systems. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. In 3D, users were less accurate identifying regions in the most complex data sets (i.e., left artery systems). Whereas in 2D, performance was the same regardless of task complexity.

For the top performing combination (2D with non-rainbow), the low ESS regions that were not identified by participants were generally the smallest in area of all the regions in a given data set. These regions were also very close to the diverging point in the color map bordering between "low ESS" and "normal". In the other conditions, there was no observed regularity in the low ESS regions missed.

We observed a negligible number of false positives (only 6 instances across all users). These false positives occurred in both color schemes, but all occurred only in 3D representations.

We additionally examined the difference in accuracy between the left and right coronary artery branches as shown in Fig. 3.8. The left branch systems are more complex due to additional bifurcations inherent to the anatomy. In our data sets, the left artery systems ranged from 7 to 10 branches (M = 8)[6] and the right artery systems ranged from 4 to 7 branches (M = 6). On average the left artery systems had 25 low ESS regions and the right artery systems 17 low ESS regions. There was no significant difference in accuracy between these data types in 2D ($Z = -1.57$, $p = 0.117$). However, in 3D participants were significantly less accurate when identifying regions in left artery systems than in right systems ($Z = -3.35$, $p = 0.001$). This provides evidence that in 3D the performance accuracy decreases with increased data complexity.

Efficiency: On average, participants spent less time per image in the 2D condition (M = 56 seconds) than in the 3D condition (M = 73 seconds) and this difference was statistically significant ($t(21) = -2.52$, $p = 0.021$)[7] (see Fig 3.9). In the 2D condition, we also observed a significant effect of color mapping on the average task completion time ($t(21) = -1.57$, $p = 0.013$): participants spent less time on images using the rainbow color map (M = 47 seconds) than on images using the diverging color map (M = 65 seconds). We saw no such effect in the 3D condition ($t(21) = -0.351$, $p = 0.741$). Even though participants completed 2D images more quickly with the rainbow color map, they had poor accuracy as described in the previous section.

Therefore, we next look at the average amount of time taken to identify a low ESS region (i.e., total time spent on an image divided by the number of low ESS

[6] M = mean.

[7] t(#) = t-test with # of participants.

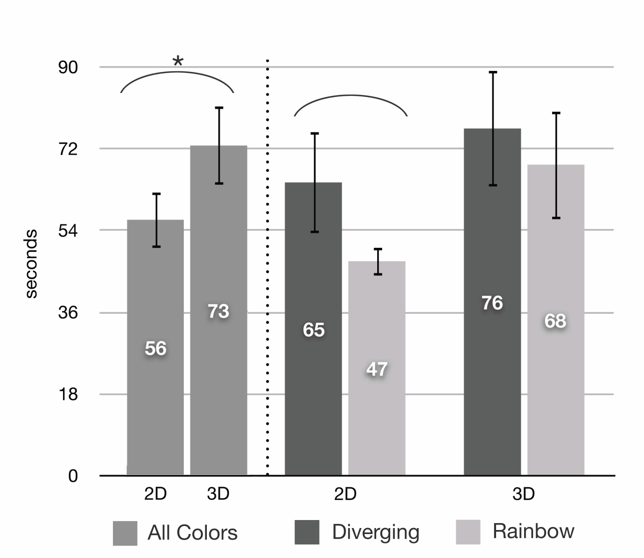

Figure 3.9: Average total time spent on each image broken down by 2D and 3D representation, and color. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. Participants completed tasks more quickly in 2D than 3D.

regions identified). As illustrated in Fig. 3.10, there is a significant difference between participants' performance in 2D and 3D ($Z = 115.5$, $p < 0.001$) with participants identifying regions more quickly in 2D (M = 2.5 seconds per region) than in 3D (M = 7.8 seconds per region). There is no significant effect of color mapping in 2D with respect to this measure ($U = 44$, $p = 0.439$, $r = 0.169$) indicating that the utility and effectiveness of the 2D representation outweighs the effect of color in regards to rate of identifying regions. However, we did observe a significant effect of color mapping in 3D ($U = 18$, $p = 0.009$, $r = 0.567$) with participants identifying regions approximately twice as fast with the diverging color map. Thus the effect of the rainbow color map on task efficiency has a greater impact in 3D than in 2D.

Subjective Responses: When examining the subjective statements, statisti-

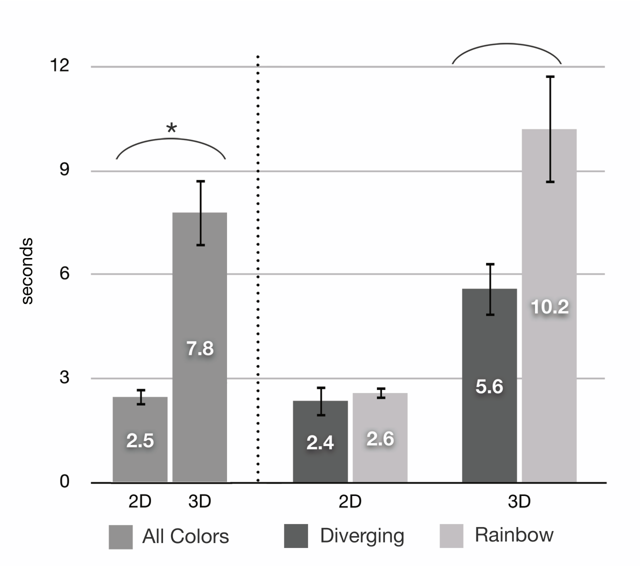

Figure 3.10: Average rates of seconds per region to identify broken down by 2D and 3D representation, and color. Error bars correspond to the standard error and the asterisks indicate results of statistical significance. Participants were more efficient in 2D, and in 3D there was a significant difference in participant performance between color schemes.

cally significant differences were observed between the 2D and 3D representations.

As shown in Fig. 3.2, on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree) participants indicated that on average it was easier to identify low ESS regions in 2D than in 3D ($Z = -75.5$, $p < 0.001$). They also reported that it was more efficient to identify regions in 2D ($Z = -72.5$, $p < 0.001$), and that they were more confident they found all the low ESS regions in 2D ($Z = -68.0$, $p < 0.001$). There was no statistically significant effect of dimensionality of presentation on participants' confidence that what they marked as low ESS were in fact really a low ESS regions ($Z = -33.5$, $p = 0.146$). This is consistent with their actual perfor-

Table 3.2: Averages of subjective responses broken down by 2D and 3D representation, and color. The four statements are rated on a 7-point Likert scale (1 = strongly disagree, 7 = strongly agree), and the asterisks indicate results of statistical significance. Participants felt it was easier and faster in 2D, and they felt more confident in 2D.

	2D	3D	* = significant	2D		* = significant	3D		* = significant
	(both color maps)			Diverging	Rainbow		Diverging	Rainbow	
Statement:									
I found it easy to identify low shear stress regions.	6.30	5.35	*	6.33	6.27		5.48	5.20	
I was able to perform the task efficiently.	6.27	5.30	*	6.21	6.33		5.55	5.03	
I am confident I found all the low shear stress regions.	6.11	5.17	*	6.30	5.90		5.55	4.77	
I am confident all the places I marked are really low shear stress.	5.84	5.63		5.91	5.77		5.85	5.40	

mance: as reported earlier, we observed very few false positives throughout the study.

We observed no statistically significant effects of color scheme on any of the participants' subjective responses. This indicates that the participants thought they did well using the rainbow color map even when in reality they did not perform as well as the participants who used the diverging color map.

3.6.6 Discussion

The results fully support our first hypothesis: participants missed fewer low ESS regions in 2D than in 3D and they completed the tasks more quickly (both in terms of total time and when comparing times spent per low ESS region identified).

This was also reflected in the verbal question portion at the end of the study sessions in which 18 out of the 21 participants said they preferred the 2D representation citing it was "easier", "more efficient", and "better for viewing the data since all the data is visible at once". Of the 3 participants who preferred the 3D representation, 2 of the participants verbally acknowledged that the 2D visualization was better and more efficient for completing the task but chose 3D as their "preferred" representation due to aesthetics.

The results also partially support our second hypothesis: in both 2D and 3D conditions participants who were presented with the data using the diverging color scheme made fewer diagnostic mistakes than those who saw the same data presented in the rainbow color scheme. The efficiency results are less equivocal: even though the rainbow color scheme resulted in faster total completion times in 2D, controlling for the number of low ESS regions identified, we saw no performance differences due to color mapping in 2D, but in 3D we observed participants being nearly twice as slow on a per region basis with the rainbow color mapping than with the diverging.

Part of the reason why the 2D representation is efficient is because people are able to easily "read" across the image and mark regions in a systematic manner. We concluded this based on the observed order in which participants identified low ESS regions and statements from participants during the verbal feedback section. In contrast, there is no obvious strategy for "reading" across the 3D representation. The 3D visualization also requires one to rotate and interact with the image, thus it takes longer for someone to view all the data. In addition to participants verbally complaining about the added interaction, participants had a difficult time remembering where they had previously identified a region of low ESS in the 3D representation. Thus in order to make the 3D visualization more effective, one would need to develop a good "mark-up" strategy such that a person knows what regions they have already identified or arteries already inspected. The 2D representation also makes it easier to identify regions of low ESS by easily exposing complex 3D features, such as artery bends and bifurcation, where low ESS regions are likely to occur. Indeed, our results demonstrated that as the complexity of the tasks increased, participants were able to maintain their accuracy in the 2D condition, but not in the 3D condition.

Additionally, based on the results of this work, our medical participants and collaborators are now convinced of the utility of a 2D data representation and appropriate color map choice:

"Three-dimensional volume visualizations provide the ability to visually follow the connections between different branches. HemoVis presents a surprisingly elegant solution to this problem in 2D by simply and cleanly plotting individual 2D "multi-spectral" presentations of all vessels concurrently, and simply superposing a graph showing their connectivity. In this manner I think the visualization is a simple yet elegant, and powerful solution for conveying a mix of innately 2D (stenosis degree) and innately 3D (endothelium) information."

"We have struggled for many years to find a way to display anatomic (i.e., geometric) data and endothelial shear stress data in a comprehensive and intuitive manner. I think HemoVis elegantly solves this problem and should be useful to clinicians and researchers alike. HemoVis is especially helpful in highlighting critical areas of low endothelial shear stress and assessing their relationship to the surrounding anatomy."

"It was surprising to find that different color mapping techniques can render the task of identifying low shear stress regions less ambivalent. By enhancing the perception of identifiable patterns in this complicated problem that spans multiple independent scientific disciplines and hence differently trained scientists, it becomes that much easier to reach significant conclusions. One can only wonder in just how many other instances we make our task more difficult than it needs to be simply by maintaining the status quo. I for one am now more open to consider visualization an integral aspect of research, particularly before dismissing hypotheses that rely on identifying complicated data patterns."

3.7 Conclusions & Future Work

Through our formative qualitative user study, we have developed a task taxonomy for blood flow visualization and we have developed a new 2D tree diagram representation of coronary artery trees. The results of our quantitative study demonstrate that the 2D representation is not sensitive to increased complexity in the task and users are more accurate and efficient at identifying regions of interest in a 2D representation than a 3D representation, and that the rainbow color map can significantly reduce a person's accuracy and efficiency.

We are continuing to develop HemoVis based on the principles and results of this study. Also, even though the 2D representation is more accurate and efficient for our tasks, having a 3D representation is still essential for surgical planning. We will investigate the most effective ways to connect these two representations through linked views in future work. We also plan to investigate other user interface designs and interactions for HemoVis. For example, if a doctor were in a clinical setting that allowed for detailed study of the data and interaction, could adding filters to narrow the range of ESS focus or adjustment of color scale parameters be useful.

The work presented in this chapter is broadly applicable to other domain applications as well as visualization in general. The new 2D tree diagram representation utilized in HemoVis is applicable to the visualization of other branched anatomical structures (e.g., cerebral and venous arterial systems, pulmonary systems) and general fluid dynamical pipe structures (e.g., engineering). In terms of general visualization, this work serves as both an example and template of how to convince users of good visualization practices. In this case, a success story of changing users' opinions with particular regard to appropriate dimensionality of data representation and color choice. This work not only shows a real world example demonstrating just how significant an impact rainbow color can have on a user's task, but also a way for other researchers to counter this issue by demonstrating to their users how color impacts their task performance.

4

Evaluation of Filesystem Provenance Visualization Tools

In the previous chapter we saw a design case study and evaluation from biomedicine. In this chapter we see an example from computer science in which visualization tools and techniques for the analysis of filesystem provenance data are developed and evaluated. In a format similar to the previous chapter, we will discuss a case study in computer science, the new visualization tool and technique developed as part of the project, and the results of a user study evaluation.

4.1 Motivation

Provenance is the history of derivation of an object. In filesystems, provenance data is a recording of the relationships of reads and writes between processes and files. In quantitative analysis of scientific data, file provenance offers many benefits. For example, a researcher may receive a third-party data set and wish to use it as a basis for further research or compare the provenance of a repeated experiment to diagnose an error. Without provenance metadata attached, they would have no record of the computations and operations that generated or manipulated that data set. File provenance also offers benefits for IT administrators. Routine administration tasks, such as analysis of log files or finding where viruses were introduced into a system, can be made more challenging by the presence of hidden dependencies. Provenance can expose these dependencies and the interwoven causes of system errors.

Because of these types of potential benefits, systems researchers predict that within the next ten years all mainstream file systems will be provenance aware [123]. However, the provenance data that existing systems generate is of only limited use. For one, the sheer amount of data recorded dwarfs a human's ability to parse through it. Provenance data can be large, sometimes as much as an order of magnitude greater than the data for which the provenance is recorded [80]. Visualization can be a powerful tool for understanding these large data sets.

Many provenance researchers use graph visualization tools to examine interrelationships on a small subset of nodes. The inability of these tools to visualize large data sets, however, limits the scale at which these data sets can be analyzed and prevents researchers from taking full advantage of the entire provenance database. For instance, a provenance-aware storage system (PASS) recording of a five-minute compilation job of the Berkeley Automator suite of tools has 46,100 nodes and 157,455 edges. Provenance data sets spanning multiple days or even months can grow dramatically in size. Examining only a small subset of the data at one time eliminates the benefits of recording such a comprehensive set of information in the first place. These forgone benefits include the ability to compare the activity of multiple process executions over time or the ability to see dependencies linking the cause of a system fault outside the expected region of error. Having an effective, scalable visualization for provenance data is crucial part of the filesystem's effectiveness as an aid for data analysis, system understanding, and knowledge discovery.

In collaboration with the PASS (Provenance-Aware Storage System) [1] group at Harvard University, we set out to develop a new visualization tool to enable easy and effective exploration of filesystem provenance data. Through a qualitative study with provenance domain experts, we put together a set of tasks to address their visualization needs and gain a better understanding of their current visualization practices. Through a task-driven iterative design process we developed a novel filesystem provenance visualization called InProv that utilizes a radial layout (Fig. 4.1, *middle* & *bottom*). The tool also incorporates our new time-based hierarchical node grouping method. This new method was inspired by feedback from our qualitative user study. The method more closely matches the user's mental model of node creation and evolution, and enables more intuitive data exploration. In-

[1]http://www.eecs.harvard.edu/syrah/pass/

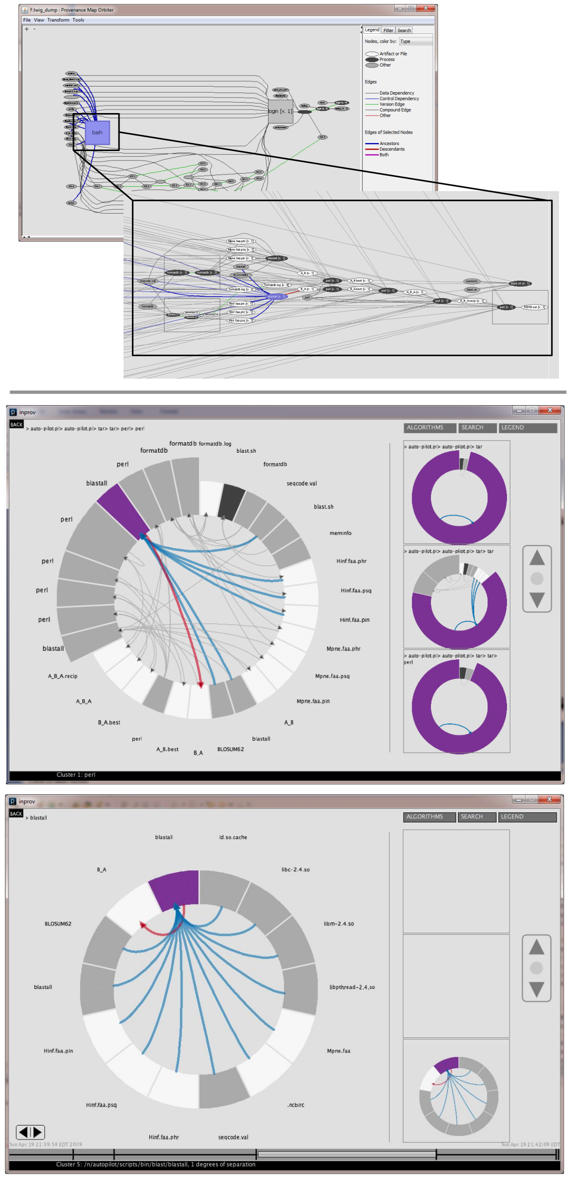

Figure 4.1: Top: A screenshot of Orbiter, a conventional node-link visualization tool for filesystem provenance data, displaying a data set with the process tree node grouping method. A zoom-in on one of the square "super nodes" in Orbiter reveals the sub-nodes and their connections to other nodes. **Middle:** Screenshot of InProv, our new radial-based visualization tool for browsing filesystem provenance data, displaying the same data with the same node grouping as *top*. **Bottom:** Screenshot of InProv with our new time-based node grouping method with the same data as displayed in the other screenshots (top, middle).

Prov displays a filesystem provenance graph in a visual format conducive to exploration in addition to focused querying. The current design and implementation of InProv has been tested on graphs of up to 60,000 nodes.

To evaluate the effectiveness of InProv with its radial layout compared to Orbiter [112], a conventional node-link diagram (Fig. 4.1, *top*), we designed and performed a quantitative user study. The study also compared the effectiveness of our new time-based hierarchical node grouping method to a conventional method. The user study was a mixed between and within-subject user study and evaluated each tool with several real world example data sets. Domain experts knowledgable in the topics of our sample data were recruited to participate in the study. The results of the study demonstrate that the new time-based hierarchical node grouping method is more effective for analyzing data in both tools, and that InProv is more accurate and efficient than Orbiter for analyzing large complex data.

The first contribution of this chapter is a set of requirements for filesystem provenance data analysis based on our interviews with domain experts. Our second contribution is InProv, a new radial layout visualization tool for browsing filesystem provenance data. Our third contribution, developed to make InProv more effective by identifying the most important nodes and processes in a system, is a new time-based hierarchical node grouping method for provenance data. Our final contribution is the results of our quantitative user study. We present statistically significant results that people are more accurate and efficient using our new time-based node grouping method, and that the radial based visualization tool, InProv, is more accurate and efficient than Orbiter at analyzing large complex data. Subjectively participants found InProv to be easier to use and preferable to Orbiter. Our user study results also demonstrates one of the first examples of gender differences in visualization tool performance.

4.2 Related Work

Provenance Data Visualization: The conventional visual encodings for provenance data are derived from the fields of network and graph visualization. Having effective visualizations of provenance data is necessary for a person to understand and evaluate the data [96]. The most common visualization strategy for provenance data is the node-link diagram and is employed by common provenance tools such as Haystack [74], Probe-It [40], and Orbiter [112]. With this visual encoding, nodes are represented as glyphs and edges or connections between nodes are represented as lines or curves. These tools utilize a variety of different visual encoding techniques including directed node-link diagrams [40, 74] and collapsible summary nodes [112]. A specific application area for provenance are workflows, such as visualization [79] and scientific workflows (e.g., tracking where data sets originated and how they have been manipulated). Visualizations for scientific workflows are also focused on node-link diagrams and include such tools as VisTrails [14, 33, 155] and ZOOM UserViews [22]. Unfortunately, these node-link visualization strategies are difficult to scale to provenance data sets beyond a few hundred nodes. Traditional node-link diagrams can easily become too visually cluttered for the multi-thousand node filesystem provenance data limiting a user's ability to thoroughly analyze and explore the data. In our tool, we employ an alternative radial layout with hierarchical encoding with an easily navigable time dimension to reduce visual clutter and bring the most important nodes to the forefront.

Network & Tree Diagrams: There has been extensive work in the network visualization community on effective techniques for generating and drawing large complex networks [1, 6–8, 52, 69, 136, 180]. There has also been work on the effective display of networks that change over time, usually employing animation [50, 95]. Most provenance data have hierarchical properties or attributes. Thus, we found visual encoding techniques from the tree visualization community to be useful points of reference [156]. For example, TreePlus is an example of a tree-inspired graph visualization tool that prioritizes node readability and layout stability [102]. The visual interface displays a tree, starting from the graph root or a user-specified starting node. This technique is more effective than a traditional node-link diagram for exploring subgraphs and providing "local overviews," but fails to provide a high-level overview of the relationships in the overall graph. Another tree-inspired visualization tool is TreeNetViz, which displays tree-structured network data using a radial, space-filling layout with edge bundling [55]. For large complex provenance data sets, the strategies employed by TreeNetViz, in which sectors expand, will become visually complex and is not necessarily an efficient use of screen real estate. In our work we employ a similar radial layout to TreeNetViz in which our tool expands sectors, but they expand into a full new radial plot to maximize label readability and take advantage of available screen space.

Radial Plots: Radial or circular layouts bring visual focus to the relationships between nodes rather than the relative spatial locations of nodes. One of the earliest examples of radial layout visualization was proposed by Salton et al. [154] for visualizing text data. Since then, many successful visualization tools using this radial layout have been produced to visualize everything from file systems to social network data to genomics data [41, 51, 78, 94, 108, 110, 119, 167]. Spatial encoding can reflect useful attributes for smaller graphs [19, 111], because the human eye is acutely attuned to deciphering 2D spatial positions. We employ a radial plot layout to reduce visual clutter and easily show connections and nodes relevant to our user base. Processes and unique activity are accentuated while system libraries and ubiquitous workflows such as system boot-up are minimized.

In the following sections we present a more detailed background on provenance and related terminology, discuss the domain specific set of tasks that motivated the design of InProv, and present the design and implementation of InProv. We then describe a new time-based hierarchical grouping method for provenance data developed for InProv. Finally, we present the results of our quantitative user study to evaluate the performance of InProv relative to Orbiter [112], a conventional node-link graph visualization tool. We conclude by discussing the results presented in this chapter and highlighting areas of future work.

4.3 Provenance Data

We focus on filesystem provenance data (i.e., the relationship between files and processes and their interactions). Filesystem provenance data are inherently an annotated directed acyclic graph. We tested InProv on output from PASS, a "provenance-aware storage system" created by the Systems Research Group at Harvard (SYRAH) [2]. **Nodes** may be processes (an instance of an execution of a program that may read from and/or write to files or pass data or signals to other processes), files (static representations of data), pipes (communication channels between processes), non-provenance files (files whose actions are not recorded), or "other" (filetypes unrecognized by the PASS system). **Edges** represent the

[2]http://www.eecs.harvard.edu/ syrah/

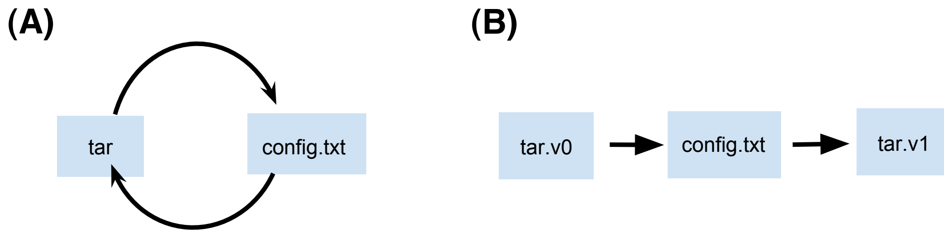

Figure 4.2: (A) There exists a cycle between tar and config.txt. (B) By versioning the tar process, PASS ensures that the graph remains acyclic.

dependency relationships between the nodes. For example, edges could represent "a process writes to a file", "a process reads from a file", "a process spawns another process", or "a user controls a process."

Each node may have a variety of attributes such as node name, filesystem path, and process node ID. This information is important to investigate specific processes or gain a deeper understanding of what is occurring in the system. Nodes also have an indegree and an outdegree. Indegree and outdegree refer to the number of edges that lead into or out of a given node.

To ensure the resulting provenance graph is acyclic, the PASS system uses a cycle reduction algorithm that assigns a version number to each node. (Fig. 4.2). The PASS system records a **timestamp** called "freezetime" as an attribute of when each versioned node is created. It also records the "exectime" when a process executes.

4.4 Requirements Analysis

We conducted an informal qualitative formative user study with provenance researchers who have been developing and using provenance capture systems for over five years. Our goal was to identify the domain specific analytic tasks that an effective visualization should address. We conducted semi-structured interviews with seven provenance data experts, all of whom work with filesystem provenance data, to learn about their data analysis and exploration tasks, current visualization solutions they use, and the limitations of their existing visualizations and workflow. The interviews lasted approximately one hour. The interviews were contextual and, in addition to answering the interviewer's questions, the interviewee demonstrated the workflow and analysis tools they were currently using. Each interviewee was asked questions relating to their current area of research, the analysis tools they used, the data formats with which they interacted, and the analysis tasks they performed.

We used affinity diagraming [21] to analyze the data from the interviews to identify common domain tasks. Despite the range of task requirements, a common theme emerged: while researchers could effectively analyze small subsets of a provenance graph, understanding the system as a whole usually required line-by-line analysis of the original (raw) data. The lack of an effective way to visualize large graphs prevented researchers from extracting an informative whole-graph analysis. We thus concluded that the ability to provide a quick summary of the overall unique system activity was a key priority. Other task requirements closely echo many canonical information visualization data exploration tasks [159].

InProv was designed to handle the following domain tasks (with analytic tasks, using definitions from Amar et al. [2], in parentheses):

1. Summarize system activity — Hierarchically group provenance graph by time of system activity (Cluster, Find Anomalies). A researcher frequently needs to analyze a provenance data set generated by someone else or a personal data set that was generated long ago. Understand such data sets requires that user quickly obtain a high-level overview or summary of the activity represented by the data set. A good visualization should highlight the main events that occurred during the recording of the data.

2. **View filtered subset of system data** — Display selected provenance subgraph (Filter). Users also frequently need to more deeply analyze a subset of a data set. For example, after obtaining a high level overview as in Task 1, a user will frequently identify one or more high level tasks that warrant more detailed analysis. Alternately, a user analyzing a current trace might already have identified objects or processes of interest and may want to view the subset of the data set pertaining to them. These are both domain-specific instances of the more general ”zoom and filter” operations. An effective visualization should allow the user to naturally select a subset of nodes, either manually (e.g., by clicking) or formally specified (e.g., using a query or filter). Although interested in only a subset of the data, most users want to view and understand these subsets in the context of the entire data set. In other words, when examining some subset of nodes in a provenance graph, the user should see the selected subset of nodes in the context of the whole graph.

3. **View node attribute details** — Display attribute value (Retrieve Value). Each node in a provenance graph typically has a variety of attributes (e.g., date created, date modified, number of dependencies, etc.). Users often wish to analyze how these metrics vary across and reflect the structure of the graph. Important metrics should be visually encoded or at least displayed in a node detail view.

4. **Examine object history** — Display provenance subgraph within one edge of queried node (Filter). The most common provenance query is the lineage query, whose response explains how an object came to be in its present state. These lineages can be quite large, depending on how long the system has been running

and/or how deep in a derivation tree the object appears. Thus, a visualization should offer a node-specific view with information on how that node was created and modified over time. This task is equivalent to a query asking for information on the ancestors of a particular node.

4.5 Time-based Hierarchical Grouping

Due to the size, scale, and varying levels of granularity of provenance data, a hierarchical grouping of the nodes in the provenance graph is necessary to ensure users can comprehend a typical data set.

We initially chose Markov Chain Clustering (MCL) [178] to cluster the provenance graph. The algorithm runs by simulating a random walk on the graph. Since nodes in the same cluster have a high probability of being connected, and two nodes in different clusters have a low probability of having an edge between them, a random path beginning in one cluster has a high probability of remaining in that cluster. If a cluster is particularly large, contained nodes were divided hierarchically into subgroups by file path because files within the same folder tend to be associated with similar workflows.

However, our initial attempts to use MCL proved ineffective. The structure of the created summary nodes did not properly communicate what was going on in the system, and the visualization's users struggled to find a way to describe the contained activity. Tellingly, one of the expert users did not even recognize that the data displayed was one of his/her own provenance data files. Furthermore, users noticed that, regardless of the data they examined, the details they could see pertained to system boot-up. This ubiquitous system boot-up activity was not pertinent to their investigations and tasks.

To have the node grouping more closely reflect the mental model of the users, we developed a time-based hierarchical grouping method that revolves around the temporal attributes of the provenance data. Through our discussions with experts in our qualitative formative user study, it became evident that understanding the filesystem provenance data was easier in many cases with a temporal context as compared to other grouping methodologies. For example, with a temporal context, a researcher can follow the exact steps a computer user took to preform a specific tasks or execute a series of programs; this provides the researcher with additional insight as to the purpose of each action.

Each job or execution in a computer system produces a burst of system activity and the recording of multiple "freezetime" and "exectime" timestamps (Sec. 4.3). These bursts of activity are usually separated by longer periods of relative inactivity. Thus, grouping together provenance nodes with roughly simultaneous timestamps allows for a hierarchical subdivision of system activity at varying levels of granularity. Hadlak et. al similarly use time attributes of data to visualize hierarchies [56]. The summarizations created by our algorithm map to the summaries of system activity provided by provenance experts (Task 1, Sec. 4.4). Feedback from users indicated this clustering approach more closely matches the users' mental models of the organization of the data (i.e., processes relevant or related to each other in a temporal context are visually near each other). This was the motivation for one of our main hypotheses in our quantitative user study (H4, Sec. 4.7).

The method we developed works as follows (Fig. 4.3): First, all the timestamps in a given set of nodes and edges are sorted chronologically. Next the "average first difference," i.e., the total duration of activity in the data set divided by the total number of timestamps, is computed. Then the timestamps are scanned in

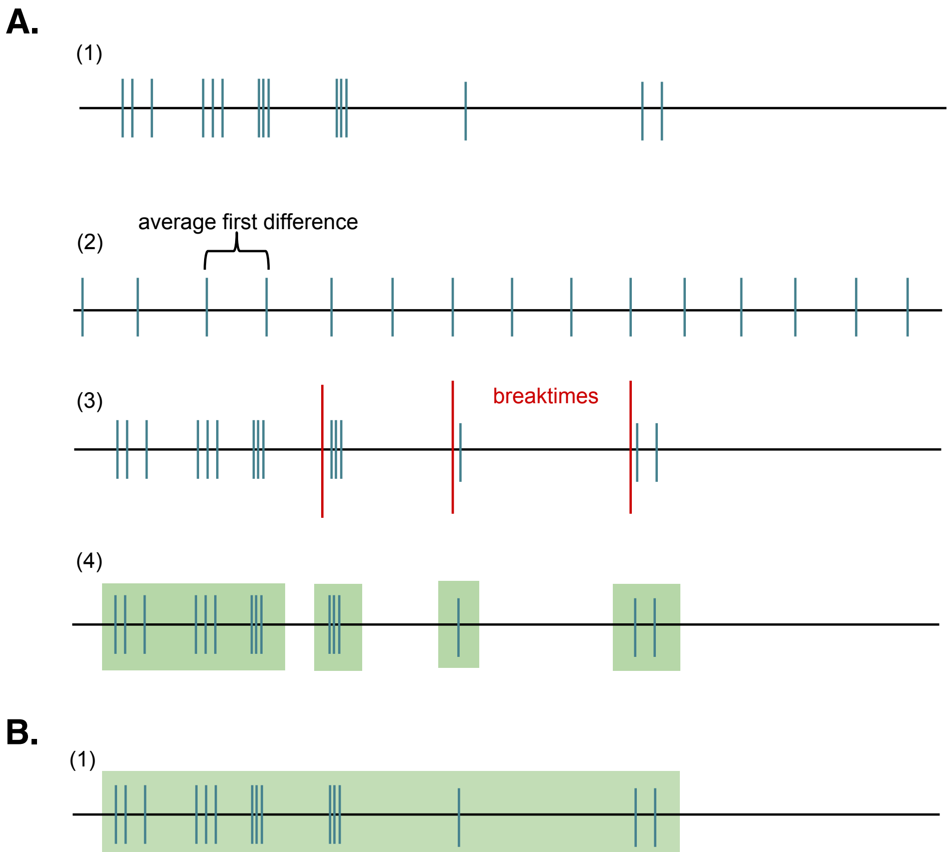

Figure 4.3: A. Time-based hierarchical grouping sorts the provenance graph according to time attributes of nodes. (1) Most system activity is distributed unevenly on a timeline. (2) Our algorithm computes the "average first difference" of timestamps, i.e., the difference between timestamps if they followed a perfectly even distribution. (3) Gaps in activity of above-average duration are marked as "breaktimes," or borders between clusters. (4) These breaktimes bookend each time-based cluster. **B.** Conventional methods group all the nodes across time into a single group based on process ID.

order and the first difference (the previous timestamp subtracted from the current timestamp) between each is computed. Whenever the first difference is above a threshold, i.e., there is a significantly long gap in recorded activity (default being twice the average first difference based on expert input and pilot testing of different thresholds), that time is recorded as a break between node groups. Nodes with activity occurring between two subsequent break times are defined as new groups.

The algorithm tries to produce between five and sixty groups, with each group limited to fifty nodes. Based on our formative study, these heuristics marked the observed limits of a user's ability to comprehend and to explore a data set. If a group has more than fifty nodes, the algorithm will attempt to divide it hierarchically into subgroups of nodes so that the user is not overwhelmed by the display of too many nodes. This hierarchical subgrouping of nodes based on time is beneficial to both "bushy" and "deep" provenance trees. Bushy trees result from widely used tools (i.e., compiler has lots of descendants) and deep trees result from continued data derivation (i.e., extract items, analyze them, re-do analysis and repeat). In both cases subdividing and grouping by temporal information will usually broaden deep trees and summarize bushy trees for easier comprehension.

One of the limitations of the current implementation is that during dense periods of activity an excessive number of nodes will be grouped at one particular time step. The other limitation is that certain patterns of user activity are sometimes not optimally split. For example, a script that compiles a tool and then immediately runs a workload that uses it. A user would expect that the compile would be in one group and the workload in another. However the workload may instead be split so that one group represents the compile plus the beginning of the workload, and the other cluster has the rest of the workload.

We plan to implement in future work "smarter" breaks in system activity (e.g., [11]). It should also be noted that this grouping method collapses versions resulting in a non-directed acyclic graph. This does not conflict with the tasks discussed in Sec. 4.4, but needs to be examined in future work if ordering is important to the task at hand.

4.6 InProv Browser

Based on our formative interviews and task-driven iterative design process with domain experts, we developed a new provenance data browser called InProv (Figs. 4.1 & 4.4). Motivated by Task 1 (Sec. 4.4), the need to have an effective high-level overview of the system, we adopted a hierarchical radial layout for the visual display of the provenance node graph as this provides focus on the overall structure of the graph and makes it easy to read the edges connecting nodes. We will show the utility for specific features of the layout in the remainder of this section. Also motivated by Task 1 (Sec. 4.4), the default node grouping method for the provenance graph in InProv is our new time-based hierarchical method (Sec. 4.5). The timeline at the bottom of the screen provides temporal context for each group. Each of these groups of nodes is displayed in the center of the screen as a ring divided into multiple sectors. Each sector in a ring is either a single node or a subgroup of nodes, visually encoded as a thicker sector (e.g., Fig. 4.1, *middle*), which can be expanded into a ring of its own (Task 2, Sec. 4.4). A text path at the top of the screen, as well as the "context view" rings on the right of the screen, provides context on the sequence of node or node group expansions. InProv was implemented using Java and Processing. We plan to make it open-source

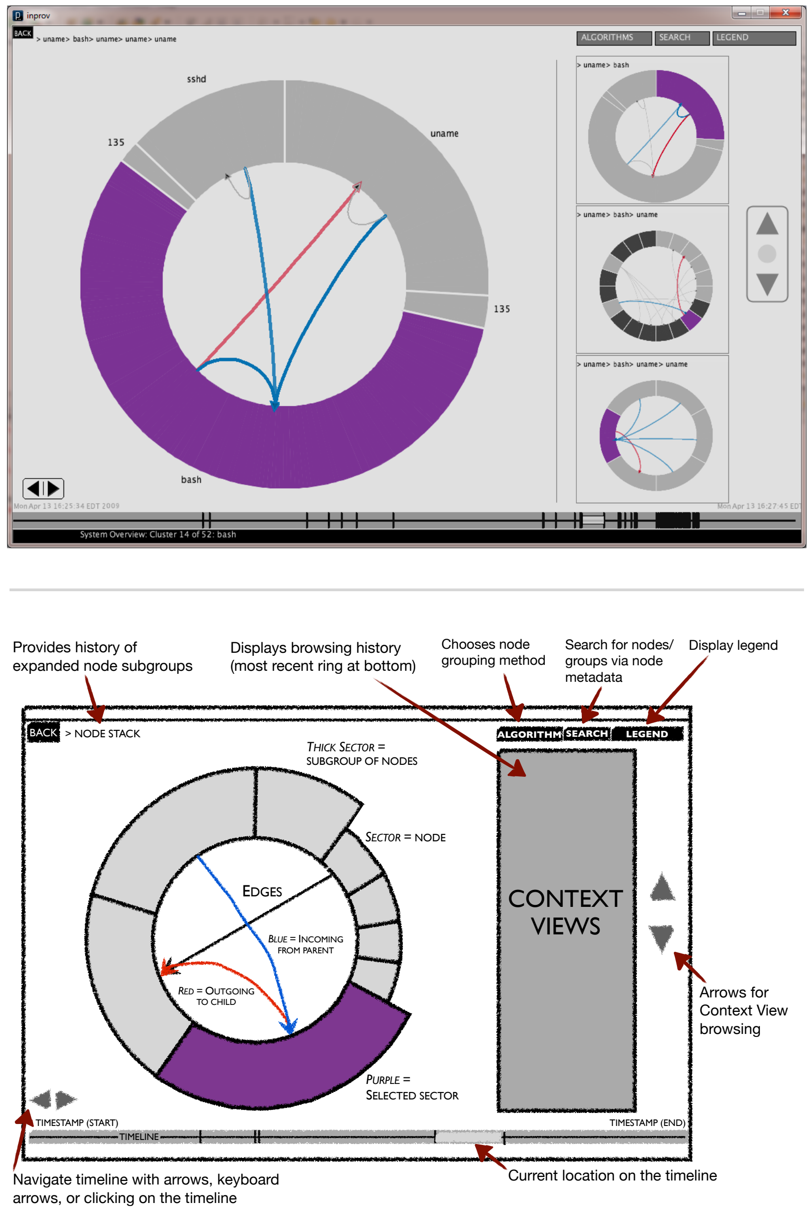

Figure 4.4: Top: Screenshot of InProv showing the interactions of the node "bash" with its parent and child nodes. The blue edges represent incoming edges from parent nodes, and the red edges represent outgoing edges to child nodes. **Bottom:** Schematic drawing displaying the key visual encodings and interaction features for InProv. The "node stack" and "context views" both provide context of browsing history as well as location within the hierarchical structure.

available.

Nodes: Nodes, visually encoded as sectors in a ring, are colored according to their type: processes are dark grey, files are white, and all other files (including non-provenance files and node groups) are grey. Subgroups of nodes are represented as thicker sectors than individual nodes (e.g., Fig. 4.1, *middle*). The width of a node subgroup sector in radians is proportional to the number of nodes it contains (e.g., Fig. 4.4, "bash" contains more nodes than "sshd" thus it covers a larger fraction of the radial plot). Nodes are drawn clockwise around the ring in order of increasing Provenance Node ID, or PNODE (analogous to the INODE of a file). InProv originally did not have a deterministic algorithm to order sectors. This was confusing to users because the same ring could look different upon multiple viewings. PNODE was chosen as an ordering index because PNODEs are assigned by the PASS system in monotonically increasing order, thus a PNODE number is an effective heuristic for creation date. This enables a "clock" metaphor, where a user can read the procession of nodes around the circle as the progression in time of node creation. To adapt InProv to display provenance information of a different format, PNODE could be replaced with any other ordinal metric, such as creation time or last modification time. This representation of ordered nodes, or groups of nodes, provides a compact easy to see representation of the system activity (Task 1, Sec. 4.4).

Edges: Edges, visually encoded as lines, are drawn in the center of the ring in the direction of data flow (i.e., from parent nodes to their children). As compared to other visual encodings, such as node-link diagrams, the radial layout's edges are clean and easy to read with minimal visual clutter (Task 1, Sec. 4.4). While canonical provenance direction flows from children to parents, following an object's history up through the chain of ancestry, this directionality was found to be counter-intuitive by participants in our formative qualitative study (Sec. 4.4), thus InProv draws edges from parent to child nodes. Edges are also drawn for subgroups of nodes. If subgroups A and B are sectors in the same ring, and a node in group A has an edge to a node in group B, an edge will be drawn from sector A to sector B (e.g., Fig. 4.4, at least one node in the "uname" group has an edge to a node in the "bash" group, but no nodes in the "uname" group are connected to any nodes within "sshd"). For more detail about the edges to and from particular sectors, a user can click and select those sectors. The incoming and outgoing edges will be highlighted with bright colors so that they visually pop from the other edges in the ring. Incoming edges, from parents, are colored blue (e.g., from "sshd" to "bash" in Fig 4.4), while outgoing edges, to children, are colored red (e.g., from "bash" to "uname" in Fig 4.4). We initially drew the edges as thin solid lines. We changed the design to arrows because edge directionality was important to users. The opacity of edges between sectors indicates how many edges there are between the two sectors. Stronger connections are more opaque and more visible. This draws the user's eye to more active connections (Task 4, Sec. 4.4).

The visualization does not distinguish between control dependency (exchanged signals), data dependency (exchanged data), or version edges (connecting different instances of the same node). The provenance researchers we interviewed explained that they did not need to distinguish these edge types for any of their primary tasks (Sec. 4.4). Since this visualization was designed to give a high level overview of a provenance data set without overwhelming the user, this design choice is reasonable.

Timeline: Each ring represents a group of system activity that happened around the same time. However, users need to be able to examine the evolution of the system over time (Task 4, Sec. 4.4), thus InProv has the ability to browse data over time. The duration of this activity is shown on the timeline (e.g., bottom of Fig. 4.4). The dates above the timeline show the earliest and latest timestamps in the data file. From these timestamps, the user can infer the duration of data collection. The duration of the currently viewed cluster is represented on the timeline as a grey rectangle. As the user scrolls left and right through the available clusters by using the left and right arrow keys or clicking the onscreen arrows, the grey rectangle moves along the timeline to update the user on his/her current contextual location. Clicking a sector will highlight its associated timestamps on the timeline as black hashmarks. The timeline partially solves the need for context by showing how the viewed cluster and any selected sectors relate to the overall graph (Task 2, Sec. 4.4). The timeline is only enabled when the data are grouped with the time-based hierarchical node grouping algorithm.

Algorithms: In addition to our new time-based node grouping method, InProv can also group nodes using a conventional "process tree" node grouping method based on control flow information [112]. This method creates summary nodes by treating processes as primary nodes and constructing a summary node for each primary node. It arranges these summary nodes in a way that reflects the process tree reconstructed from the control flow information found in the provenance metadata (Fig. 4.3, *B*). Each summary node contains a primary node and all of its immediate ancestral and descendant secondary nodes (non-processes). InProv is able to group the nodes and draw the ring(s) with either algorithm; by hovering over the "Algorithms" button, the user can choose between the time and process tree node grouping methods.

Navigation and Interaction: Hovering the mouse over a sector displays a tool tip with more information about that particular sector. This design feature was motivated by the users' need to investigate more detailed information about a particular node (Task 3, Sec. 4.4). If the sector is a subgroup of nodes, hovering will display information such as the number of contained sectors, as well as the numbers of contained files and processes. Clicking on a sector selects it, turning it purple, and clicking again on the selected sector expands it. If the sector represents a subgroup of nodes, those nodes will expand to fill a new ring (Task 2, Sec. 4.4). We investigated expanding sectors in place, as in TreeNetViz, but decided that limiting the total number of sectors displayed to the user at any given time for comprehensibility was a greater priority [55]. If the sector is a single node, the new ring will display all nodes one edge away from the current node regardless of what timestamp they were in originally. The user can thus see what connections a node has outside of the group it which it was initially displayed in (Task 4, Sec. 4.4).

Node Stack: Each time a sector is expanded, its name is added to a list of expanded node groups, or nodes, displayed at the top of the screen as a text path. Next to the "node stack" text path is a "BACK" button for returning to the previous ring (e.g., top of Fig. 4.4). This list of sectors communicates the path the user took to get to the current view. We added this feature in response to user feedback. During qualitative feedback sessions with an early version of InProv, users repeatedly complained that, upon expanding a node, they were confused as to how they had ended up in their new location and were unclear on the current view's location in the overall graph. The addition of the node stack greatly helped the users to keep contex and understand the hierarchical structure as node subgroups were expanded (Tasks 1 & 2, Sec. 4.4).

Context Views: Each time a sector is expanded, a miniature version of its previous ring and its node stack path are added to the "context view" displayed on the right side of the screen. The context view displays three rings at a time. The rings are stored starting from the bottom of the screen, where the most current ring is displayed. The context view scroll, i.e., the up and down arrows to the right of the context view, allows the user to view their navigation history. The sector that was clicked-on for expansion is colored purple in each of the context view rings. This helps the user remember their browsing history as well as give hierarchical context. For example, expanding a series of node subgroups in a ring will show the hierarchical context of the data (Task 2, Sec. 4.4). When the data is clustered by time, each break in time (as denoted by the hashmarks) has its own context view. Thus, the user's context view is not lost during navigation.

4.7 Quantitative User Study

We conducted a quantitative user study to evaluate the accuracy and efficiency of InProv compared to Orbiter, a conventional filesystem provenance data visualization tool using node link diagrams. In the same study, we also compared our new time-based hierarchical grouping method (see Sec. 4.5) to a conventional process ID node grouping method. We implemented both new and conventional node grouping methods into InProv and Orbiter for the user study.

To ensure broad relevance of the results, we included two different types of tasks, two levels of task difficulty, and four different user populations.

4.7.1 Hypotheses

Our hypotheses entering the user study were:

H.1 Participants will be able to complete tasks more accurately in InProv than Orbiter. The radial layout utilized in InProv more concisely summarizes and presents the information to users compared to the node-link diagram utilized in Orbiter. This simpler representation will enable users to more accurately complete tasks.

H.2 Participants will be able to complete tasks more efficiently in InProv than in Orbiter. Navigation and context viewing in InProv allows users to track their visited paths more easily than in Orbiter. The increased amount of zoom in or out required to explore the node-link diagram in Orbiter will make it more difficult for users to remember their visited paths.

H.3 Participants will subjectively prefer using InProv to Orbiter and find the tool easier to use. Following the reasoning in H1 and H2, users will find InProv overall easier to use for task completion.

H.4 Participants will perform tasks more accurately and more efficiently in both tools when the nodes are grouped according to our new time-based hierarchical node grouping. We hypothesized that the time-based grouping of nodes would be more consistent with the users' mental models of the historical file system activity than the hierarchal dependency grouping, thus users will be more accurate and efficient in both tools when completing tasks with the time-based grouping.

4.7.2 Participants and Apparatus

Because our use case scenarios focused on both IT professionals and scientific applications, we recruited study participants from these fields. Twenty-seven members of the Harvard community participated in the study (20 men, 7 women; 19–59 years old, M=34). Thirteen participants were professional IT staff. Ten were scientists representing domains covered by our tasks (6 bio/medical and 4 astrophysics computational scientists). The remaining 4 participants were provenance research experts. Participants received monetary compensation for their time.

We required that all participants be familiar with Linux/Unix operating systems as the minimal background knowledge required to participate in the study. We also required that all participants have normal color vision (i.e., are not color blind).

All of the user study sessions were conducted in the same indoor room utilizing the identical Lenovo ThinkPad 15" (1600x900 screen resolution) laptop running Windows Vista with Logitech wireless mouse with scroll wheel. Camtasia Studio 8 was used for screen and audio capture.

4.7.3 Tasks

We had two types of tasks. The first type was focused on finding an explicit file or process node, and the second type was focused on understanding larger concepts demonstrated by the sample provenance data. This first task type is derived from the second and fourth task requirements in our set of tasks, and the second task type is derived from the first task requirement in our set of tasks (see Sec. 4.4). The following question is an example of the first task type: "A radiologist is analyzing a patient's medical imaging data. Which process is responsible for aligning and warping the images?". The following question is an example of the second task type: “A user is complaining about their computer acting weird. Looking at the user’s provenance data from before the complaint, what was the application the user invoked?”. For each task in the study, a data set was loaded into the tool and the participants were asked a question prompting them to complete one of these two types of tasks. Participants were presented with an equal number of both task types during the study. For each task type, we had 5 instances. Out of all 10 instances, 5 of them were easy (42-346 nodes) and 5 were hard (1192-5480 nodes). The boundary between easy and difficult tasks was determined in a pilot experiment in which tasks with 10s, 100s, 1000s, and 10,000s of nodes were compared.

The tasks used real world sample data and the questions were designed to mimic such real world scenarios. The sample questions above are examples of a bio/medical imaging scenario, and an IT scenario, respectively. The wording of the questions relating to our scientific scenarios were derived from the questions asked as part of the First and Third Provenance Challenges [121, 162]. The data sets from these two challenges were used as the domain scientific data in our study. The data are standardized and publicly available [3]. The 1st Provenance Challenge’s data is on brain atlases from the fMRI Data Center and the 3rd Provenance Challenge’s data set on the Pan-STARRS project. The other IT related questions, as well as the PASS team’s sample data from these provenance challenges, are also publicly available online through the PASS Team Website [4]. All participants were presented with the same set of tasks which included tasks from multiple domains.

[3] http://twiki.ipaw.info/bin/view/Challenge/WebHome

[4] http://www.eecs.harvard.edu/syrah/pass/traces/

4.7.4 Procedure

Each study session started off with a basic demographic survey and a series of multiple choice questions to assess each participant's prior knowledge of Linux/Unix operating systems as well as filesystem provenance. Next, the participants were presented with two pages of background information on filesystem provenance data in order to make sure all participants possessed a basic understanding of provenance. Then the participants received instruction (demonstrated and read from a script by the experimenter) on how to use each of the two visualization tools and received a practice task to perform with each tool. The practice tasks were similar to the tasks given during the main study. The practice data sets also were of varying difficulty (one "easy" and one "hard"), thus representative of the two levels of complexity in data they would see during the study. Finally, the participants moved on to the main part of the study and completed 8 tasks alternating between tools for each task.

For the main part of the experiment, participants were given a series of eight tasks with a specific data set associated with each. All participants completed the same set of mixed-domain tasks with identical associated data, and task orderings were balanced both in the order of tool presentation as well as difficulty level. Participants alternated between the two tools for each task in order to minimize learning effects. The participants also alternated between pairs of "easy" and "hard" data sets. Genders and populations (i.e., astronomer, bio/medical scientist, IT specialist, and provenance expert) were balanced between the two algorithms, between which tool they started with, and between which data difficulty they started with.

The participants were instructed to "talk out loud" while completing the tasks, to verbally state when they had a preliminary guess, and to state what their final answer was. This additional verbalized information was critical to evaluating the participant's performance. The verbalization, applied to a relatively simple task with static data, and was applied equally in all conditions to all participants. The duration of each task was timed from the screen capture from the moment the participant first moved the mouse (after they finished reading the question) to the statement of their final answer. Except for the practice tasks, users were not given feedback during the session whether their answer was correct or incorrect.

With both tools, the participants were given complete freedom to highlight/select nodes, pan/browse the visual representation, zoom in/out, and expand node groups. The terminology, color encodings and node labels were identical in both tools' UIs. To advance to the next level of the hierarchy in a node subgroup, users double-clicked a thick subgroup sector in InProv while in Orbiter users could either zoom in with the scroll wheel on the mouse or double-click on a "summary node" box. When using Orbiter, users could pan around the node-link diagram by clicking and dragging. (No panning is required with the radial layout of InProv.) When viewing data with the time-based hierarchical grouping algorithm, both tools would display a timeline along the bottom of the screen and a user could either click the left-and-right arrows with a mouse, use the left and right arrow keys on the keyboard, or click/drag the timeline marker to navigate.

The study participants were asked to complete each task in as timely a manner as possible. If the participant was unable to complete the task within 5 minutes, the participant was asked whether he or she had a final answer and was given the post-task questionnaire. Based on a pre-study pilot, it was observed that if a participant was not able to provide an answer within 5 minutes then the participant generally was never able to provide the correct answer.

After each task was completed, the participants were presented with a questionnaire with nine questions to respond to on a 7-point Likert scale. The first six questions were the raw NASA-TLX standard questions for task load evaluation [61, 62], and the remaining three questions gauged subjective ease of use, self-efficacy, and subjective assessment of the tool's effectiveness for the task: "How easy was it to use the tool?", "How confident are you in your answers(s)?", and "How easily were you able to accomplish this task?".

At the end of the session, participants were verbally asked which visualization tool they preferred to use and why, and whether they had any other general comments or feedback. The entire session lasted approximately 60 minutes.

4.7.5 Experimental Design & Analysis

The study was a 2 x 2 x 2 mixed between- and within-subject design with the following factors and levels:

- Tool (InProv or Orbiter)
- Difficulty (size, complexity) of data (easy or hard)
- Node grouping method (process tree or time-based)

Tool and difficulty were within-subject factors and node grouping method was a between-subject factor. Our dependent measures were number of correctly completed tasks, time to complete a task, and participants' subjective responses recorded on a 7-point Likert scale. Accuracy was a binary measure (i.e., correct or incorrect answer), and the answer keys for each data set were generated by filesystem provenance data experts.

Because many participants waited until the five minute time out to declare their answer, the timing data had a bimodal distribution and we thus used a non-parametric test to analyze them. Also, because normal distributions cannot be assumed for Likert scale responses, we used non-parametric tests to analyze subjective responses as well. For within-subjects comparisons (i.e., to investigate the effects of tool and difficulty) we used the Wilcoxon signed rank test, and for between-subjects comparisons (for investigating the effects of node grouping method) we used the Mann-Whitney U test.

For accuracy, we used a Generalized Linear Model with a binomial distribution. In the model we included the following factors and interactions: tool, data difficulty, node grouping method, tool×difficulty, and tool×node grouping. Additionally, we controlled for effects of population (astronomy, bio/medical, IT, provenance) by including it as an additional factor. Finally, we also included gender and gender×tool as additional factors because our initial analyses revealed possible gender-related differences in performance.

4.8 User Study Results

4.8.1 Accuracy

We observed a significant main effect of node grouping method on accuracy with participants being more accurate with the new time-based hierarchical node grouping as compared to the process tree node grouping method ($\chi^2_{(1,N=216)} = 22.74$, $p < 0.001$) as shown in Fig. 4.5.

Participants were on average more accurate using InProv (M=73%) than using Orbiter (M=67%), but the difference was not statistically significant ($\chi^2_{(1,N=216)} = 2.000$, $p > 0.05$). As we expected to potentially see a difference in performance be-

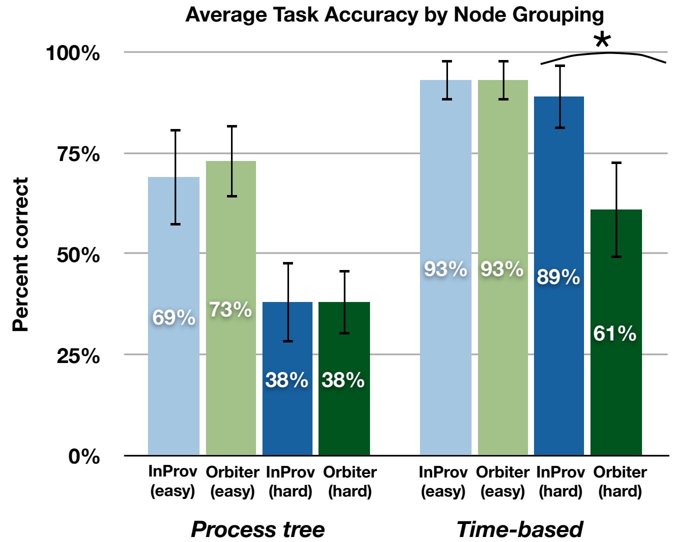

Figure 4.5: Average accuracies of participants sorted by difficulty level, tool, and node grouping method. Error bars correspond to the standard error and the asterisks indicate results of statistical significance.

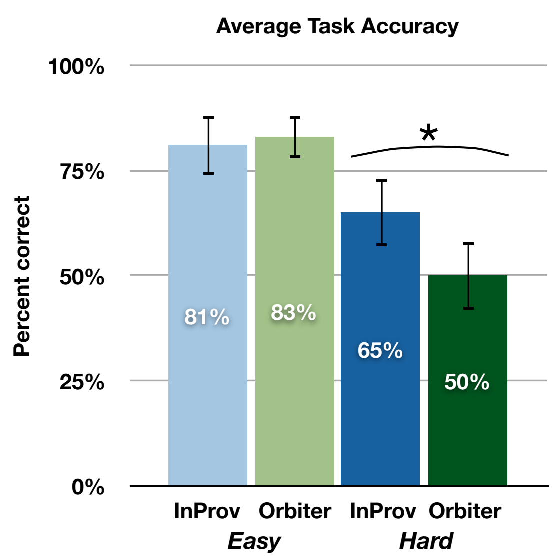

Figure 4.6: Average accuracies of participants sorted by data difficulty level (easy vs. hard) and tool. Although performance was comparable between tools for easy data, InProv had higher accuracy for hard data.

tween easy and hard data sets, as it has been observed that node-link diagrams are difficult to read if too dense [53], we repeated the analysis separately for the two difficulty levels. While there were no significant effects of tool on performance for easy data sets ($\chi^2_{(1,N=108)}$ = 0.861, p = 0.354), on hard data sets participants were significantly more accurate with InProv than with Orbiter ($\chi^2_{(1,N=108)}$ = 7.787, p = 0.005). These results are illustrated in Fig. 4.6.

4.8.2 Efficiency

As shown in Fig. 4.7, there was a main effect of node grouping method on average completion time (U = 30, p = 0.003, r = -0.570). With both tools, participants were almost twice as efficient with the time-based node grouping method as compared to the process tree method.

Participants were more efficient on average with InProv than with Orbiter, but the difference was not significant when both data set difficulty levels were considered together (z = -1.201, $p > 0.05$). Breaking down the analysis by difficulty, there was no observed effect of tool with easy data but there was a statistically significant effect with hard data (z = -2.057, p = 0.040). As shown in Fig. 4.8, participants were more efficient with an average task completion time for hard data of 167 seconds with InProv compared to 206 seconds with Orbiter.

4.8.3 Subjective Responses

We observed statistically significant effects of tool and node grouping method on participants' responses to certain subjective questions as shown in Table 4.1. As discussed in Sec. 4.7.4, the participants rated their answers on a 7-point Likert

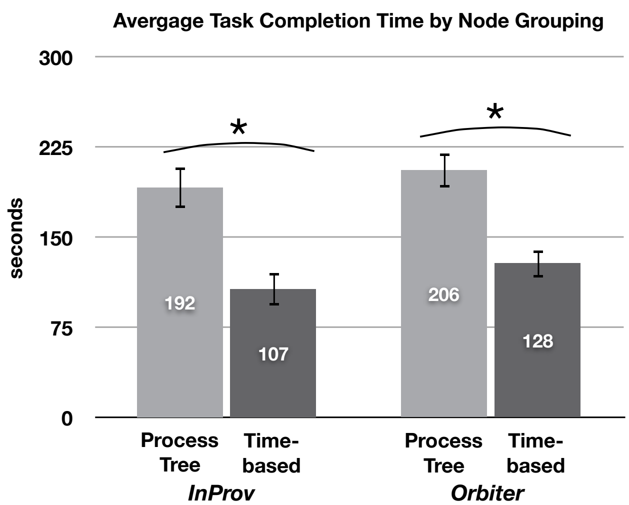

Figure 4.7: Average task completion time for each tool broken down by node grouping method. Participants were more efficient using the time-based method. Error bars correspond to the standard error and the asterisks indicate results of statistical significance.

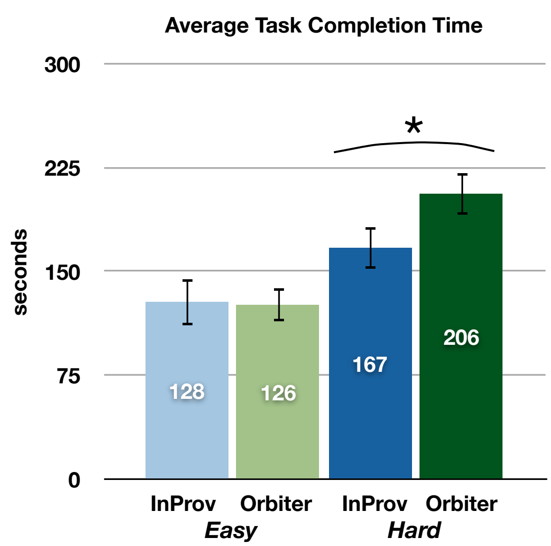

Figure 4.8: Average task completion time for easy and hard data sorted by tool. Participants in the study took longer to complete hard data tasks with Orbiter. Error bars correspond to the standard error and the asterisk indicates results of statistical significance.

Table 4.1: Average subjective data responses to the raw NASA-TLX and subjective questions. The answers were rated on a 7-point Likert scale. Asterisks indicate results of statistical significance, and "d" is the Cohen's d effect size.

#	Question	*Tool*				*Node grouping method*			
		InProv	**Orbiter**	**Sig.?**	**d**	**Process tree**	**Time-based**	**Sig.?**	**d**
1	How much mental and perceptual activity was required? Was the task easy or demanding, simple or complex? *(1 = low, 7 = high)*	**3.2**	3.6	*	0.30	3.5	3.3		
2	How much physical activity was required? Was the task easy or demanding, slack or strenuous? *(1 = low, 7 = high)*	**2.0**	2.3	*	0.18	**1.7**	2.6	*	0.62
3	How much time pressure did you feel due to the pace at which the tasks or task elements occurred? Was the pace slow or rapid? *(1 = low, 7 = high)*	3.2	3.5			3.9	**2.8**	*	0.64
4	How successful were you in performing the task? How satisfied were you with your performance? *(1 = low, 7 = high)*	4.8	4.4			3.8	**5.3**	*	0.94
5	How hard did you have to work (mentally and physically) to accomplish your level of performance? *(1 = easy, 7 = hard)*	**3.1**	3.5	*	0.25	3.4	3.2		
6	How irritated, stressed, and annoyed versus content, relaxed, and complacent did you feel during the task? *(1 = relaxed, 7 = stressed)*	**2.9**	3.3	*	0.20	3.4	2.8		
7	How easy was it to use the tool? *(1 = easy, 7 = hard)*	3.4	3.7			3.4	3.6		
8	How confident are you in your answers(s)? *(1 = low, 7 = high)*	4.6	4.4			3.7	**5.3**	*	0.93
9	How easily were you able to accomplish this task? *(1 = easy, 7 = hard)*	3.7	4.0			4.4	**3.3**	*	0.76

scale with questions 1-6 being raw NASA-TLX measures. These measures positively reflect upon InProv with participants stating it required less mental activity (Q1), less physical activity (Q2), required less work (Q5), and was less stressful (Q6).

A statistically significant effect of node grouping method was also evident in the subjective data. These measures positively reflect upon the new time-based node grouping method with participants stating they felt less time pressure (Q3), were more successful performing their task (Q4), were more confident (Q8), and found it easier to accomplish their task (Q9). The one measure which favored the process tree node grouping is that participants stated it required less physical activity (Q2). These statistically significant results for node grouping method also have strong effect sizes (Table 4.1).

Finally, as part of the qualitative feedback solicited at the end of the study sessions, participants were asked which tool they preferred using. Participants overall preferred InProv (56%) to Orbiter (41%), with one participant stating that he/she preferred "neither." The reasons most commonly cited by those who preferred InProv include that it was "easier to navigate", "easier to see the data", and "looks nicer". Those who preferred Orbiter most commonly cited that it "has a data representation I am used to seeing" and that it is "easier to understand". Of those who preferred InProv, 67% used the time-based node grouping method in the study. In contrast, of those who preferred Oribter 64% used the process tree node grouping method in the study.

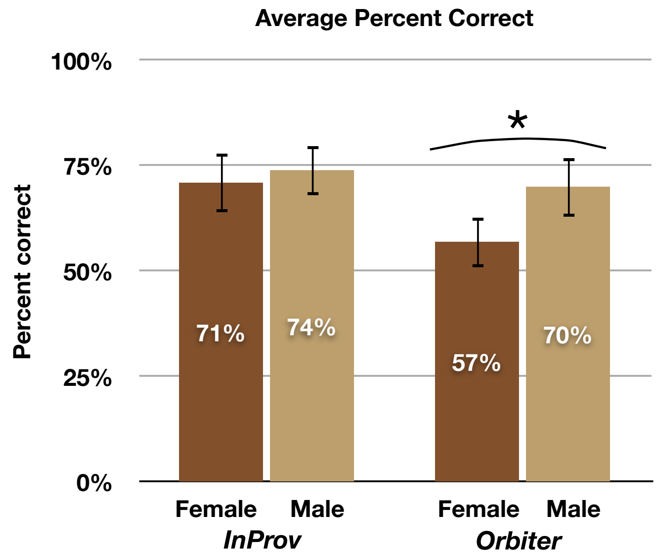

Figure 4.9: Average accuracy for each tool broken down by gender. Although men and women had comparable preformance with InProv, men were significantly more accurate than women when using Orbiter. Error bars correspond to the standard error and the asterisks indicate results of statistical significance.

4.8.4 Additional Analyses

We observed a significant interaction effect between tool and gender on accuracy for both easy ($\chi^2_{(1,N=108)}$ = 4.275, p = 0.039) and hard ($\chi^2_{(1,N=108)}$ = 6.672, p = 0.010) data. As shown in Fig. 4.9, although men and women performed similarly with InProv with 74% and 71% average accuracies, respectively, men were significantly more accurate (M=70%) than women (M=57%) when using Orbiter. The Cohen's effect size value (d = 0.36) suggests a slight to moderate practical significance for the difference in accuracy when using Orbiter.

4.9 Discussion

The results of our user study demonstrate that the time-based node grouping method resulted in significantly faster and more accurate performance with both tools than the conventional process tree method. These results provide support for our fourth hypothesis. The method's ability to pull the most relevant processes to the top of the hierarchy, and present the system boot-up processes only in the first time step or two, made a large difference in participants' performance. Through the qualitative feedback session at the end of each study, participants commented on how easy it was to use the tools with this grouping since the time element helped them understand the data and system activity. The time metaphor was more intuitive to interpret the events captured in the provenance data compared to the process tree node grouping. With the time metaphor, participants were able to easily reconstruct the original user's actions.

These strong results based on the node grouping methodology are evidence of how important it is to pick a node grouping method for network and graph based data that both presents the most relevant data to the user as well as matches the user's mental model. Regardless of visual encoding, node grouping will affect how a viewer sees, reasons through, and interprets the data presented to them in a visualization.

Our first and second hypotheses were partially supported. Although InProv did not significantly improve accuracy or efficiency for all data set difficulty levels, InProv did prove to be more accurate and more efficient when dealing with large (i.e., $>$1,000 nodes) data sets compared to Orbiter. Thus, as the data increase in size and complexity, InProv with its radial layout is able to maintain higher accuracy levels than Orbiter with its node link diagram (Fig. 4.6). In other words, task completion on large complex data sets was more accurate with the radial layout and significantly more accurate than Orbiter when utilizing the time-based node grouping method.

Examining the cases in which participants gave incorrect answers, the most common reason was because they ran out of time. Our pilot study results showed that if a participant was unable to provide an answer by the 5 minute mark, then the participant was never able to provide a correct answer. This trend was primarily seen with participants who used the process tree node grouping algorithm. The other common reason for incorrect responses was that participants would get lost in the hierarchy. This was, again, more frequently observed with the process tree node grouping method since the hierarchy was so deep compared to the time-based node grouping. Also, a common problem in Orbiter, was that participants would get lost browsing the large node-link diagram and forget where certain nodes or node clusters were located.

The results of our subjective data analysis support our third hypothesis: participants overall preferred InProv over Orbiter. The subjective ratings reveal that participants overall found InProv easier to use, requiring less work, and with the hard data tasks they felt more successful and more confident in their answers. However, there was a slight trend with node grouping method in which those who preferred InProv had the time-based node grouping method whereas those who preferred Orbiter had the process tree node grouping method. Also, those who preferred Orbiter commonly stated their preference was due to the fact that they were familiar with node-link diagrams, even if they had higher accuracy with InProv.

We also observed an unexpected effect in our study: a statistically significant interaction effect between gender and tool on accuracy. Although men and women had comparable performance with InProv, women performed worse using Orbiter with a much lower accuracy rate. However, the distribution of overall preferred tool by women matched the same response distribution as men, and women gave no verbal feedback that was significantly different from men. Women also had the same distribution of age, educational background, area of expertise, expertise with Linux, and previous knowledge of provenance data as the men in our study. Because our participants were all highly trained professionals in their respective fields, the results are unlikely to be due to differences in education or general cognitive ability. Other than gender, we are not able to find another factor that could have affected the women's performance.

We believe our results are possible evidence of gender specific differences in software design. There are known low level differences between genders that have been observed in psychology lab studies such as differences in spatial reasoning (e.g., [82, 101]). However, the question remains whether these low level gender differences can translate up to higher level tasks or interactions such as problem solving skills and, more specifically, computer visualization software. To date gender differences have been observed, for example, in confidence levels using computer software [15, 31], problem solving strategies [16], behaviors and interaction techniques with software [15, 16, 176], and hardware interfaces [39, 165].

Our observation of women having poor performance with Orbiter, within the context of this previous work, is one of the first examples of measured gender differences in visualization. Given the statistically significant systematic differences presented in our study, but with a small to moderate effect size, it seems worthwhile for future research to look deeper at potential gender-related differences in visualization and related interfaces. This could help to identify best practices in designing visualizations and interfaces for all users.

4.10 Conclusions & Future Work

We are continuing to develop InProv for filesystem provenance data exploration. Based on feedback from the quantitative evaluation, we plan to add functionality to load and view multiple files at the same time to support data set comparison. Incorporation of graph difference algorithms will help with this multi-file comparison, and incorporation of the ability to connect directly to provenance databases will enable comparisons and faster data exploration. We also plan to investigate additional or alternative methods for grouping nodes to enable more efficient or different analyses of the data such as fingerprinting-based pattern matching, manual classifications by the user, and machine learning techniques based on user-classified data sets. Finally, we plan to scale both InProv as well as the time-based node grouping method so that they will be able to handle data sets containing hundreds of thousands of nodes.

The results of the quantitative evaluation imply that radial layouts, with the right node grouping method, can be an effective visual encoding for provenance data. The visual encoding we developed in InProv may also be applicable to other types of provenance data and to network data in general. We hope that providing a tool that offers an intuitive summary of provenance data sets will help researchers and developers utilizing provenance enhanced systems, especially those dealing with large data sets. We also hope that the availability of a better tool for understanding provenance will promote the adoption of provenance recording systems and encourage more research in the provenance field.

5

What Makes a Visualization Memorable?

In this and the following chapter we will discuss the results of evaluations designed to study fundamental cognitive properties of visualizations, specifically properties of memorability. In this chapter we present the results of an experiment designed to measure the memorability of visualizations, and the observed trends of what may contribute to a visualization's inherent memorability.

5.1 Motivation

The Visualization community has recently witnessed a divide over the value and impact of excessive chart annotation and decoration (i.e., "chart junk"). The conventional view, promoted by visualization experts such as Edward Tufte and Stephen Few, holds that visualizations should not include chart junk and should show the data as clearly as possible without any distractors [44, 45, 173, 174]. This view has also been supported by psychology lab studies, which show that simple and clear visualizations are easier to understand [37, 92].

At the other end of the spectrum, researchers have published that chart junk can possibly improve retention and force a viewer to expend more cognitive effort to understand the graph, thus increasing their knowledge and understanding of the data [12, 25, 73]. However, the findings of these studies have been widely debated [44, 45].

What researchers agree on is that chart junk is not the only factor that influences how a person sees, interprets, and remembers a visualization. Other aspects of the visualization, such as graph type, color, or aesthetics, also influence a visualization's cognitive workload and retention [25, 73, 179]. To disentangle these confounding factors we set out to answer the basic question: "What makes a visualization memorable?" Clearly, a more memorable visualization is not necessarily a more comprehensible one. However, knowing what makes a visualization memorable is a step towards answering higher level questions like "What makes a visualization engaging?" or "What makes a visualization effective?". Recent work has shown that memorability of images of natural scenes is consistent across people, suggesting that some images are intrinsically more memorable than others, independent of an individual's contexts and biases [76]. We are interested in understanding if these findings hold for visualizations, and what key factors make some visualizations *intrinsically* more memorable than others.

Here, we designed and executed a study to measure the memorability of visualizations. Specifically, we studied the memorability of visualizations as images to better understand their intrinsic memorability. While we did not specifically study the memorability or comprehensibility of the underlying data presented in the visualization in the current work, identifying which type of visual information is *memorable* or *forgettable* provides a basis for understanding a number of cognitive aspects of visualizations. This is because given limited cognitive resources and time to process novel information, capitalizing on memorable displays is an effective strategy. Research in cognitive psychology has shown that conceptual knowledge is an organizing principle for the storage and retrieval of information in memory. For instance, details of a story or a picture that are consistent within an existing schema are more likely to be remembered than those that are not [4, 88]. Recent large-scale visual memory work has shown that existing categorical knowledge supports memorability for item-specific details [27, 88, 89]. In other words, many additional visual details of the image come for free when retrieving memorable items. Understanding the memorability of visualizations provides a baseline for leveraging these cognitive capabilities.

For our research, we first built a new broad taxonomy of static visualizations that covers the large variety of visualizations used across social and scientific domains. These *visualization types* range from area charts, bar charts, line graphs, and maps to diagrams, point plots, and tables. Next, we scraped over 5,693 real world visualizations from a variety of websites covering different areas of visualization publications (e.g., news media, scientific journals, infographic blogs, etc.).

We present a breakdown of visualization types by publication sources, showing some interesting visualization strategies and biases. Based on the distribution of visualization types "in the wild" we took a representative sample of 2,070 single-panel visualizations from our database and annotated them with certain *visual attributes* that we consider to be informative for memorability, such as the data-ink ratio and the visual density. We then used these 2,070 visualizations in an online memorability study we launched via Amazon's Mechanical Turk with 261 participants. This study allowed us to gather memorability scores for hundreds of these visualizations, and determine which visualization types and attributes were more memorable. While previous experiments have demonstrated that some visualizations are easier to remember than others, this is the first study that systematically analyzes this intuition. We believe this opens a new domain of investigation at the interface between human cognition and visualization design.

5.2 Related Work

Perception Theory and the Chart Junk Debate: Researchers have explored the perception of individual graph types based on tasks and data encodings [37, 92, 132]. More recently, there have been a number of studies aiming to evaluate the impact of embellishments on visualization memorability and comprehension [12, 23, 25, 44, 73, 113, 179]. Bateman et al. conducted a study to test the comprehension and recall of graphs using an embellished version and a plain version of each graph [12]. They showed that the embellished graphs outperformed the plain graphs with respect to recall, and the embellished versions were no less effective for comprehension than the plain versions. There has been some support for the comprehension results from a neurobiological standpoint, as it has been hypothesized that adding "visual difficulties" may enhance comprehension by a viewer [25, 73]. Other studies have shown that the effects of stylistic choices and visual metaphors may not have such a significant effect on perception and comprehension [23, 179]. While there have been studies evaluating memorability and perception of graphical layouts for specific types of graphs, such as the work by Marriott et al. for network diagrams [113], there has not yet been a memorability study to target a wide variety of visualizations.

In response to the Bateman study, Stephen Few wrote a comprehensive critique of their methodology [44], most of which also applies to other studies. A number of these studies were conducted with a limited number of participants and target visualizations. Moreover, in some studies the visualization targets were designed by the experimenters, introducing inherent biases and over-simplifications [12, 25, 179]. We reduced our biases by compiling a large database of thousands of real-world visualizations and enrolling a large and more diverse set of participants on Amazon's Mechanical Turk. And while previous studies confound perception, recall, comprehension, and measurements of insight, we focus purely on memorability of the visualizations as images to remove any obfuscation by other variables.

Visualization Taxonomies: Within the academic visualization community there have been many approaches to creating visualization taxonomies. Traditionally many visualization taxonomies have been based on graphical perception models, the visual and organizational layout, as well as the graphical data encodings [19, 42, 145, 159]. Our proposed taxonomy most closely aligns with this approach. However, in existing taxonomies, statistical charts are often considered as a group, even though they cover a broad range of visualization types. We propose a taxonomy with distinct categories for statistical charts based upon the visual encodings of data and the elementary perceptual tasks enabled by them. Our taxonomy also includes newer visualization types, such as text and matrix visualizations, which do not appear in previously published taxonomies.

Another approach to visualization taxonomies is based on the underlying algorithms of the visualization and not the data itself [144, 169]. There is also recent work on taxonomies for interactive visualizations and the additional tasks they enable [67, 68, 145, 159]. Both of these approaches are not applicable in our case since we focus on a large number of static visualizations for which we do not have algorithm or task classifications.

Outside of the academic community there is a thriving interest in visualization collections for the general public. For example, the Periodic Table for Management [104] present a classification of visualizations with a multitude of illustrated diagrams for business. The online community Visualizing.org introduces an eight-category taxonomy to organize the projects hosted on their site [103]. InfoDesignPatterns.com classifies visualization design patterns based upon visual representation and user interaction [17]. Our taxonomy is more comprehensive and identifies a dozen main graph types with many subtypes that span the variety of visualizations we found online. We were inspired by the reference guide by Harris [59], who provides a comprehensive reference for graphic representations, but no taxonomical classification.

Cognitive Psychology: In our study we apply techniques from previous work in the visual cognition community on evaluating the memorability of natural images of objects and scenes [27, 76, 77]. These studies have demonstrated that the differences in the memorability of different images are consistent across observers, which implies that memorability is an intrinsic property of an image [76, 77]. Brady et al. [27] tested the long-term memory capacity for storing details by detecting repeat object images when shown pairs of objects, one old and one new. They found that participants were accurate in detecting repeats with minimal false alarms, indicating that human visual memory has a higher storage capacity for minute details than was previously thought.

More recently, Isola et al. have annotated natural images with attributes, measured memorability, and performed feature selection, showing that certain features are good indicators of memorability [76, 77]. Memorability was measured by launching a "Memory Game" on Amazon Mechanical Turk, in which participants were presented with a sequence of images and instructed to press a key when they saw a repeat image in the sequence. The results showed that there was consistency across the different participants, and that people and human-scale objects in the images contribute positively to the memorability of scenes. That work also showed that unusual layouts and aesthetic beauty were not overall associated with high memorability across a dataset of everyday photos [76].

In our study we apply the same methods of measuring memorability to visualizations. In contrast to the prior work that focused on natural images and real-world objects, visualizations are artificial representations of data. Our study contributes not only to the field of visualization but also adds memorability results for artificial images to the cognitive psychology literature.

5.3 Visualization Taxonomy

In order to address the span of visualization types we found across visualization sources we created a new taxonomy for static (i.e., non-interactive) visualizations. The taxonomy classifies static visualizations according to the underlying data structures, the visual encoding of the data, and the perceptual tasks enabled by these encodings. It contains twelve main visualization categories and several popular sub-types for each category. In addition, we supply a set of properties that aid in the characterization of the visualizations. This taxonomy draws from the comprehensive vocabulary of information graphics presented in Harris [59], the emphasis on syntactic structure and information type in graphic representation by Englehardt [42], and the results of Cleveland and McGill in understanding human graphical perception [37]. A full break-down of the taxonomy and several visual properties is shown in Table 5.1. A visual version of this table with several examples is provided in Figures 5.1 through 5.5.

The *properties* are additional visual encodings that may apply to any of the visualization categories. Each property may also have subcategories. *Dimension* represents the number of dimensions (i.e., 2D or 3D) of the visual encoding. *Multiplicity* defines whether the visualization is stand-alone (*single*) or somehow grouped with other visualizations (*multiple*). We distinguish several cases of multiple visualizations. *Grouped* means multiple overlapping/superimposed visualizations, such as grouped bar charts; *multi-panel* indicates a graphic that contains multiple related visualizations as part of a single narrative; and *combination* indicates a graph with two or more superimposed visualization categories (e.g., a line plot over a bar graph). The *pictorial* property indicates that the encoding is a pictogram (e.g., a pictorial bar chart). *Pictorial unit* means that the individual pictograms represent units of data, such as the Istotype (International System of Typographic Picture

Table 5.1: Summary of our visualization taxonomy, including visualization properties and attributes, for the memorability experiment. See Figures 5.1 through 5.5 for the full version of the taxonomy with sample figures.

CATEGORY	**SUBTYPES**
Area	Area Chart (Area Chart, Overlapped Area Chart, Stacked Area Chart); Proportional Area Chart (Aligned Area Chart, Centered Area Chart, Overlapped Area Chart, Stacked and Linked Area Chart)
Bar	Bar Chart (Bar Chart, Grouped Bar Chart, Stacked Bar Chart, Circular Bar Chart, Waterfall Chart, Bullet Graph)
Circle	Belt Chart; Donut Chart; Pie Chart; Sector Graph
Diagram	Flow Chart; Illustration or Rendering; Sankey Diagram; Timeline; Venn Diagram
Distribution	Box-and-Whisker Plot; Distribution Curve; Dot Array; Histogram; Point Graph; Stem-and-Leaf Plot; Stripe Graph; Tally Graph
Grid & Matrix	Heatmap
Line	Contour Graph; Density Graph; Line Graph (Line Graph, Circular Line Graph, Trend Line (and Residual Graph)); Slopegraph; Star Plot; Surface Graph; Vector Graph
Map	Flow Map; Geographic Map (Geographic Map, Street Map); Statistical Map (Choropleth Map, Contour Map, Distorted Map, Plotted Map)
Point	Dot Plot; Scatter Plot (Bubble Graph, Scatter Plot, (Trend Line and) Residual Graph, Trilinear Scatter Plot)
Table	Table; Text Chart
Text	Phrase Net; Word Cloud; Word Tree
Trees & Networks	Trees and Networks (Graph, Matrix Representation, Tree, Treemap); Hive Graph; Hierarchical Edge Bundling
Properties	*Dimension (2D; 3D), Multiplicity (Single; Multiple; Grouped; Multi-panel; Small Multiples; Combination), Pictorial (Pictorial; Pictorial Unit), Time (Time Series)*
Attributes	*Black & White [yes, no], Number of Distinct Colors [1, 2-6, $\geq$7], Data-Ink Ratio [good, medium, bad], Visual Density [low, medium, high], Human Recognizable Objects [yes, no], Human Depiction [yes, no]*

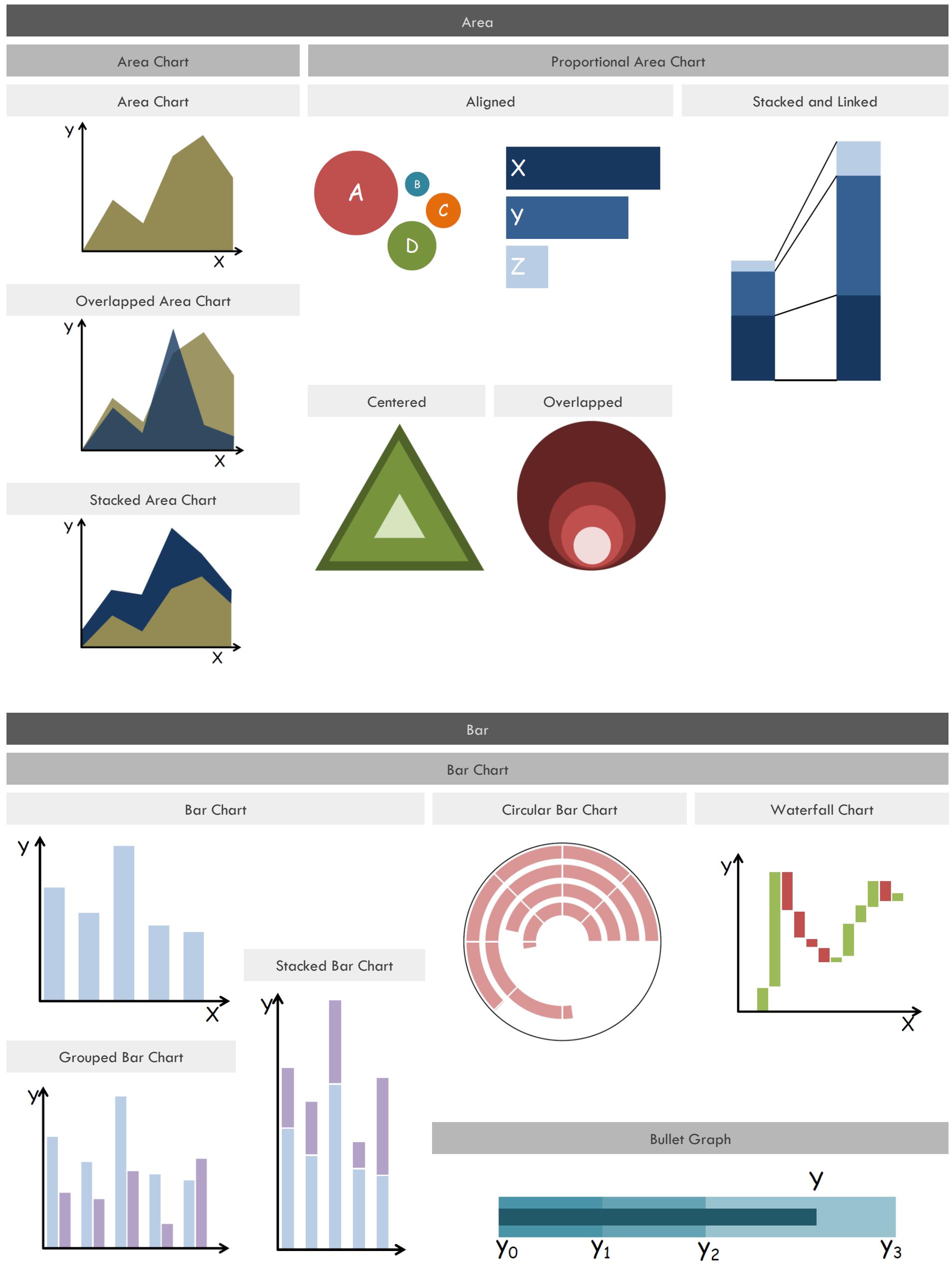

Figure 5.1: Example visual representations of the *area* and *bar* categories of our visualization taxonomy.

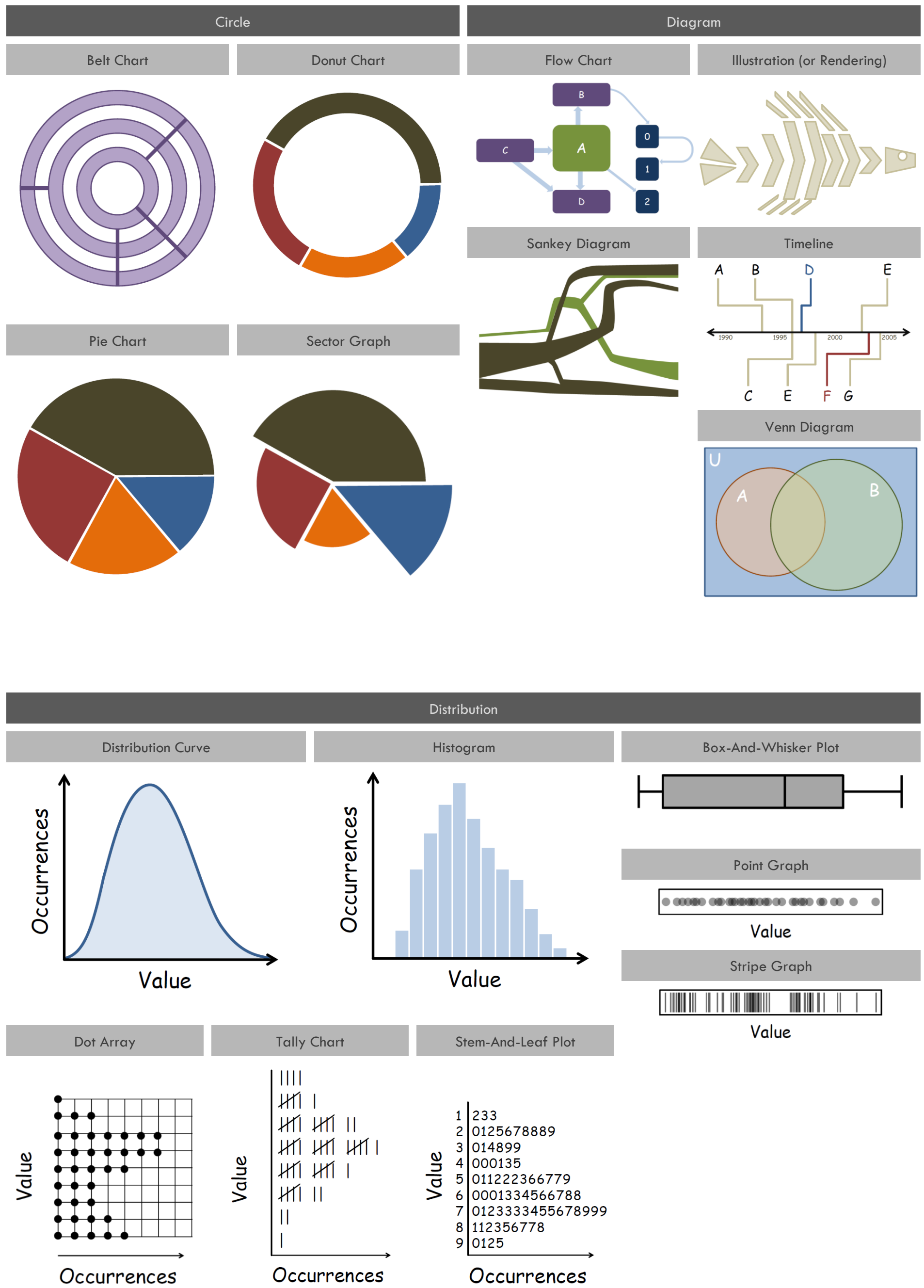

Figure 5.2: Example visual representations of the *circle*, *diagram*, and *distribution* categories of our visualization taxonomy.

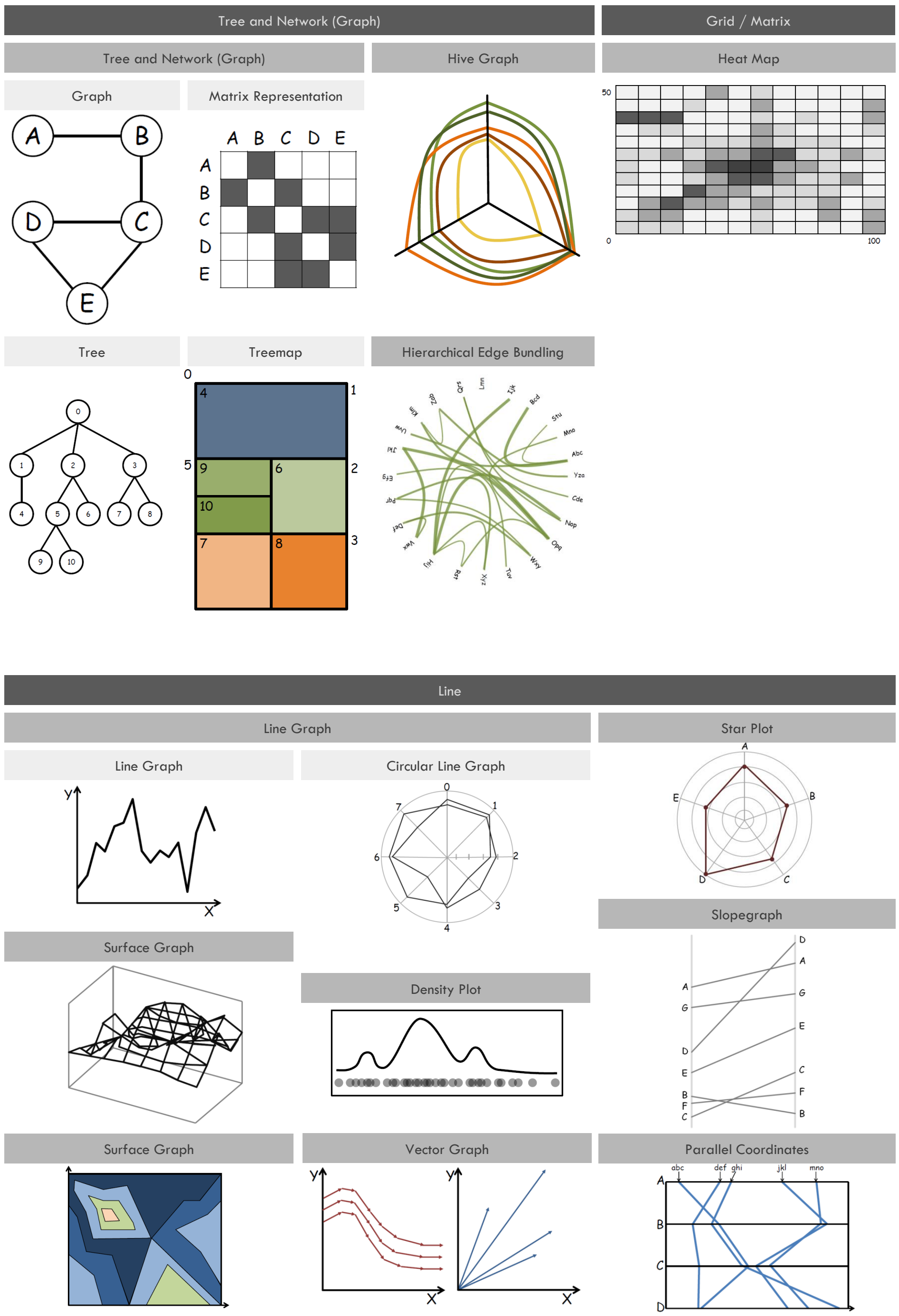

Figure 5.3: Example visual representations of the *tree and network*, *grid/matrix*, and *line* categories of our visualization taxonomy.

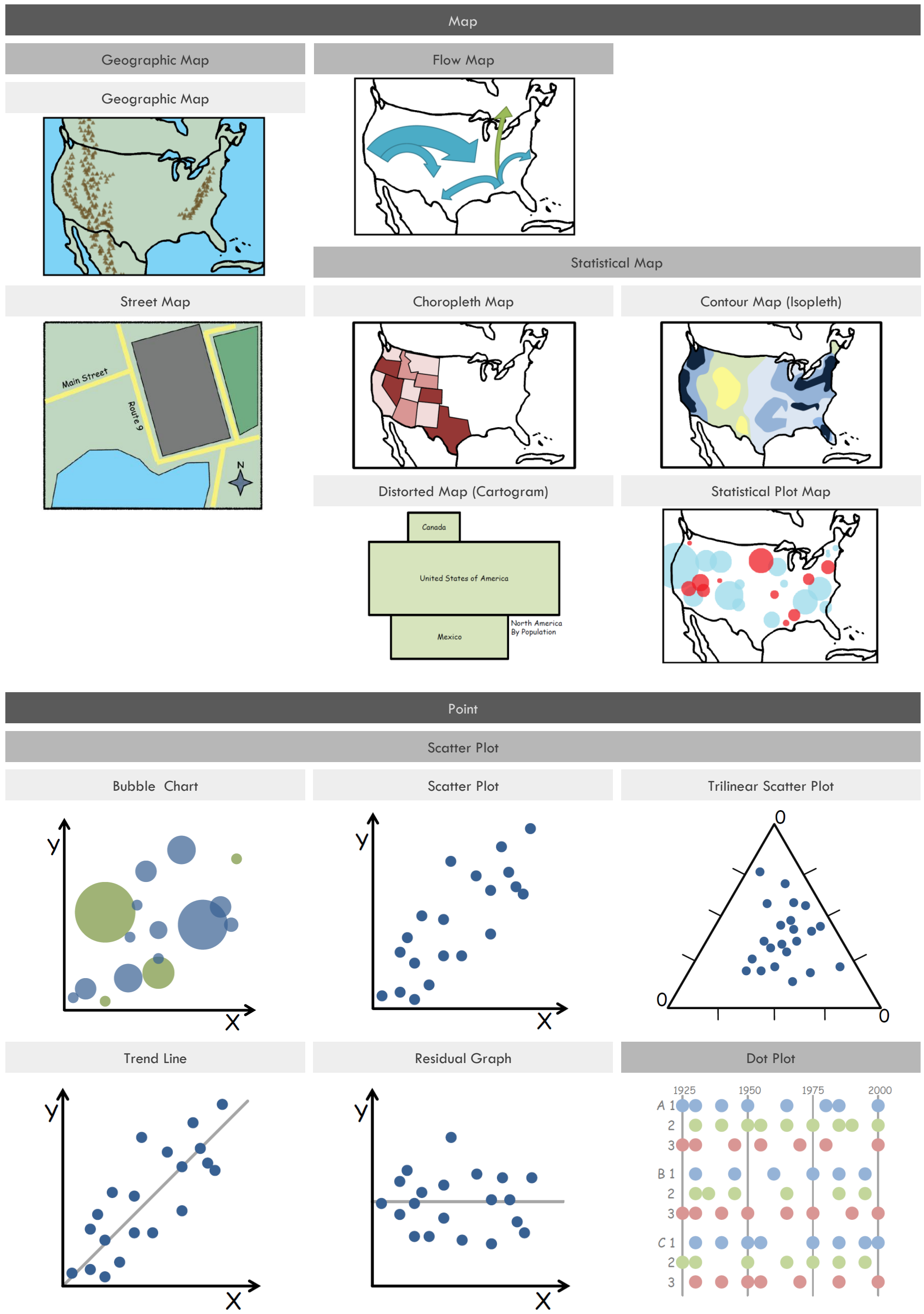

Figure 5.4: Example visual representations of the *map*, and *points* categories of our visualization taxonomy.

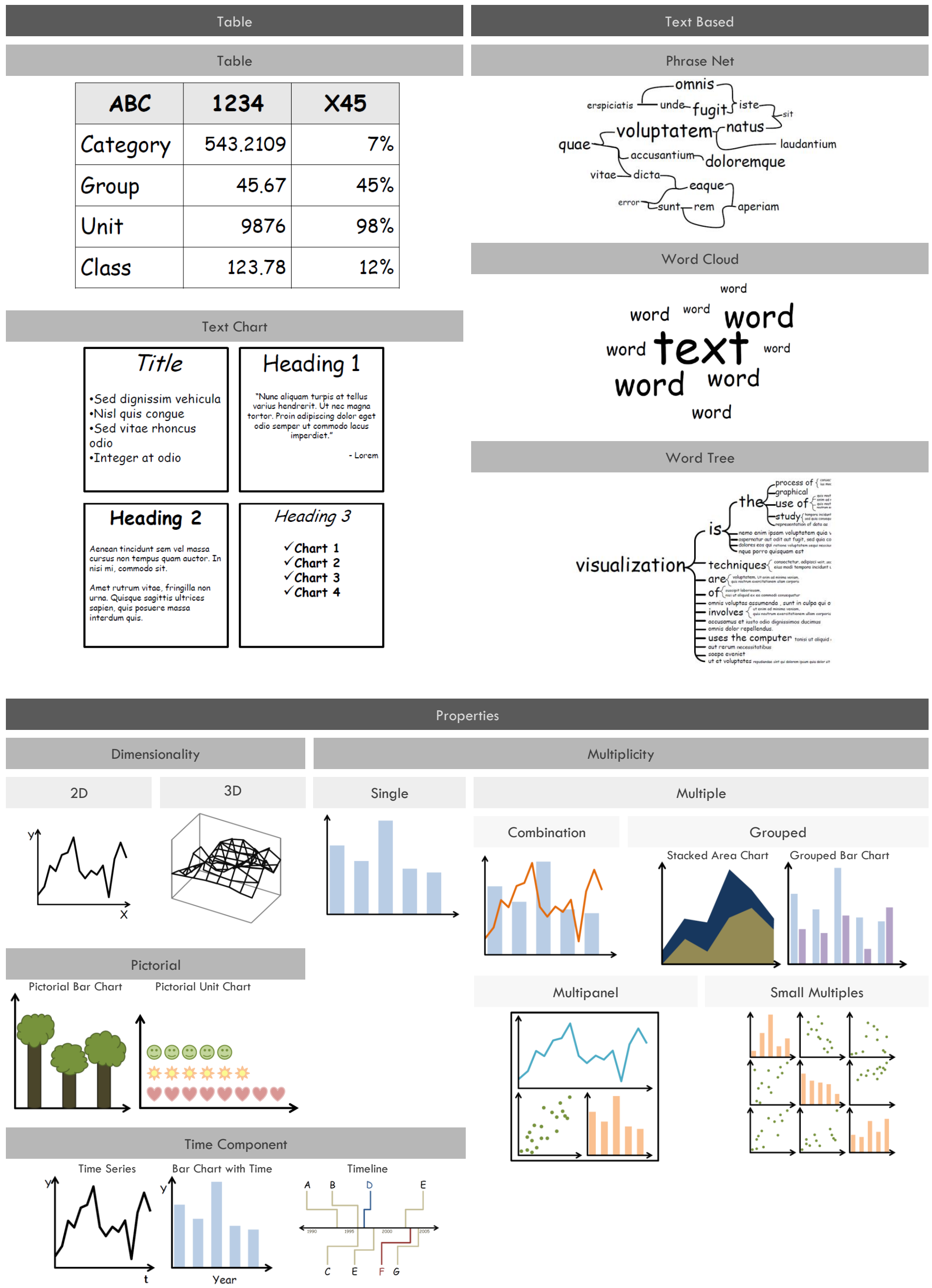

Figure 5.5: Example visual representations of the *table*, and *text based* categories of our visualization taxonomy, as well as visual depictions of the visualization properties.

Education), a form of infographics based on pictograms developed by Otto Neurath at the turn of the 19th century [126]. Finally, *time* is included, specifically as a *time series*, as it is such a common feature of visualizations and dictates specific visual encoding aspects regarding data encoding and ordering.

In order to gain insight into the effect of different visualization qualities on memorability we also defined a series of visual *attributes* that we use for analysis in our memorability experiment (Sec. 5.6). The first two attributes, *black & white* and *number of distinct colors* give a general sense of the amount of color in a visualization. A measure of chart junk and minimalism is encapsulated in Edward Tufte's *data-ink ratio* metric [174], which approximates the ratio of data to non-data elements. The *visual density* rates the overall density of visual elements in the image without distinguishing between data and non-data elements. Finally, we have two binary attributes to identify pictograms, photos, or logos: *human recognizable objects* and *human depiction*. We explicitly chose to have a separate category for human depictions due to prior research indicating that the presence of a human in a photo has a strong effect on memorability [77]. These attributes are separate from the properties in our taxonomy as they are subjective measures and are not tied to the data encodings.

5.4 Data Collection and Annotation

In order to have a large number of real world examples for our memorability experiment we scraped the web to collect 5,693 data visualizations. To ensure a breadth of visualization types, design aesthetics, and visualization domains, we focused on the visualization sources listed in Table 5.2.

These particular web sites were chosen because each contained a large number

Table 5.2: List of visualization sources, their websites, and the respective number of visualizations in the database.

Source	Total (single)	Website(s)	Per website (single)
Government / World Organizations	607 (528)	US Treasury Dept.	141 (117)
		World Health Organization (WHO)	464 (411)
News Media	1187 (704)	Wall Street Journal	609 (309)
		Economist	519 (378)
		National Post	55 (17)
Infographics	1721 (490)	Visual.ly	1721 (490)
Scientific Publications	2,178 (348)	Nature	2,178 (348)
TOTAL	5,693 (2,070)		

of static visualizations that could be automatically scraped without requiring a large manual clean-up effort. We noticed that certain visualization sources (in particular newspapers and magazines) do not provide many of their only-in-print visualizations in digital form online. Also, some websites could not be scraped as their websites were poorly structured or constantly changing in an inconsistent manner. Finally, we chose to include only one source for infographics (Visual.ly) since most infographics websites cross-post the same images, thus leading to an excessively high rate of duplicate visualizations.

All of the 5,693 visualizations were manually categorized as single or multiple. Many of the infographics visualizations were categorized as multiple and were excluded from the study. In total we identified 2,070 single visualizations, i.e., stand-alone visualizations with one panel. They were further annotated with the twelve main categories of our taxonomy (Table 5.1) plus the binary property pictorial to identify images with human-recognizable elements. The annotations were done by ten Harvard University undergraduates who had completed the Harvard introductory visualization course. The students received an introduction to the taxonomy and were monetarily compensated for their work. In the future we plan to further annotate our database with all the graph sub-types and properties of our taxonomy using Amazon's Mechanical Turk.

5.5 Analysis of Visualization Types

Our annotated data enables us to study the distribution of online visualization types across publication venues. Examining the ratio of single to multiple visualization images per source, as shown in Table 5.2, we see that visualizations categorized as multiple tend to be most numerous in scientific publications and infographic sources. These multipanel visualizations are primarily used when narrative is involved, and having multiple visualizations are necessary for, e.g., explaining a concept or telling a story. These visualizations are usually designed to stand alone, without an associated article or paper, and thus are fully responsible for telling and encompassing the whole story. There is also a very high percentage of multiple visualizations in the scientific publication category. There are two primary explanations for this observation. First, like infographics, multiple individual visualizations are combined in a single figure in order to visually explain scientific concepts or theories to the journal readers. Second, combining visualizations into a single figure (even if possibly not directly related) saves page count and money. In contrast, a very high ratio of single visualizations is seen in government / world organizations. These visualizations are usually published one-at-a-time within government reports, and there are no page limits or space issues as with scientific journals.

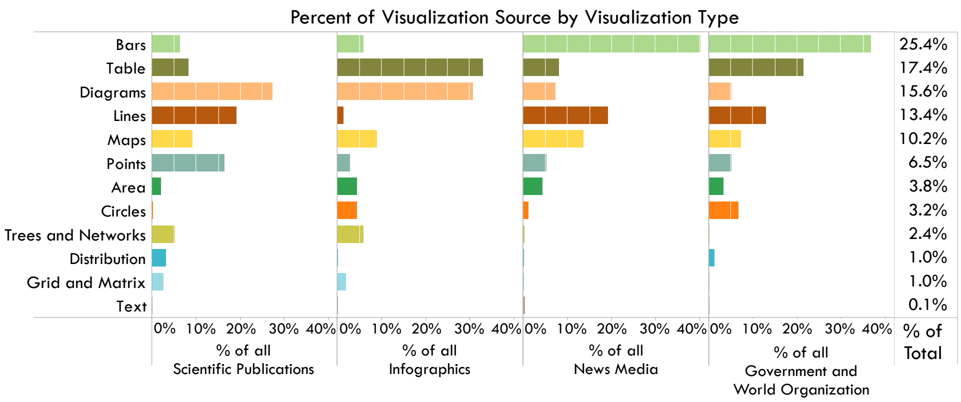

Figure 5.6: Breakdown of visualization categories by visualization sources based on 2,070 single, static visualizations.

Analyzing the single visualizations, we see distinct trends of visualization categories between visualization publication venues as shown in Fig. 5.6. Scientific publications, for example, have a large percentage of diagrams. These diagrams are primarily used to explain concepts from the article, or illustrate the results or theories. Also included are renderings (e.g., 3D molecular diagrams). The scientific articles also use many basic visual encoding techniques, such as line graphs, bar charts, and point plots. Domain-specific uses of certain visual encodings are evident, e.g., grid and matrix plots for biological heat maps, trees and networks for phylogenic trees, etc. Infographics also use a large percentage of diagrams. These diagrams primarily include flow charts and timelines. Also included in infographics is a large percentage of tables. These are commonly simple tables or ranked lists that are elaborately decorated and annotated with illustrations. Unlike the other categories, there is little use of line graphs.

In contrast to the scientific and infographic sources, the news media and government sources publish a more focused range of topics, thus employing similar visualization strategies. Both sources primarily rely on bar charts and other "common" (i.e., learnt in primary school) forms of visual encodings such as line graphs, maps, and tables. The line graphs are most commonly time series, e.g., of financial data. One of the interesting differences between the categories include the greater use of circle plots (e.g., pie charts) in government reports.

Looking at specific visualization categories, tree and network diagrams only appear in scientific and infographic publications. This is probably due to the fact that the other publication venues do not publish data that is best represented as trees or networks. Similarly, grid and matrix plots are primarily used to encode appropriate data in the scientific context. Interestingly, point plots are also primarily used in scientific publications. This may be due to either the fact that the data being visualized are indeed best visualized as point plot representations, or it could be due to domain-specific visualization conventions, e.g., in statistics.

Worth noting is the absence of text visualizations from almost all publication venues. The only examples of text based visualizations were observed in the news media. Their absence may be explained by the fact that their data, i.e., text, is not relevant to the topics published by most sources. Another possible explanation is that text visualizations are not as "main stream" in any of the visualization sources we examined as compared to other visualization types. Also worth noting is that these are observations for single, static visualizations. The distributions in Fig. 5.6 may look very different if they included, or were focused solely on, interactive or multiple visualizations.

The distribution of visualizations in this database represents a snapshot of the distributions of visualization types "in the wild." With a database of this size and breadth we can now attempt to answer the question of what makes a visualization memorable. Understanding memorability may also shed light on some of the strategies employed by the different publication venues. For example, unlike the readers of scientific journals and government reports who are already interested in the text and are pre-motivated to examine the sources, both infographic and news media outlets need to engage their audiences and capture their attention. Are they possibly employing visual strategies that help with memorability? Or are they employing strategies that they think will make a visualization memorable, but in fact are ineffective? To answer these and other questions we designed a memorability experiment discussed in the following sections.

5.6 Memorability Experiment

We ran our memorability experiment using workers on Amazon's Mechanical Turk to maintain high external validity (i.e., provide us with a diverse pool of participants for the experiment).

5.6.1 Hypotheses

Based on the authors' experience in practicing visualization, our hypotheses entering the experiment were:

H.1 Participants will perform worse (i.e., overall have a harder time remembering visualizations) as compared to natural images/photos.

H.2 A visualization is more memorable if it includes a pictogram or cartoon of a recognizable image.

H.3 A visualization is more memorable if there is more color.

H.4 A visualization is more memorable if it has low visual density.

H.5 A visualization is more memorable if it is more "minimalist" (i.e., "good" data-ink ratio).

H.6 A visualization is more memorable if it includes a "familiar" visualization type (i.e., basic graph type taught in school).

H.7 A visualization is less memorable if it comes from a scientific publication venue.

5.6.2 Target images

We selected a subset of 410 images (∼20% of the single-panel images in our database, Sec. 5.4) to be "target" visualizations. The attribute rankings (Sec. 5.3) for the target visualizations were generated by three visualization researchers. Each researcher independently rated the attributes for each visualization. In cases when all three researchers gave different rankings, the visualization was reviewed and discussed by all three researchers until a consensus was reached. If a majority of two out of three researchers agreed their ranking was applied.

Of the 410 target visualizations, 145 are extreme examples of "minimalist" (i.e., data-ink ratio = "good"), 103 are extreme examples of "chart junk" (i.e., data-ink ratio = "bad"), and the other 162 are in-between on the spectrum (i.e., data-ink ratio = "medium"). Choosing them this way allows us to measure the effects of chart junk, among other attributes, without introducing bias. The target visualizations were also chosen to match the distribution of original visualization sources as well as the distribution of visualization categories of the total 2,070 single-panel visualization population (Fig. 5.6). Thus the target population is representative of the observed diversity of real-world visualization types.

5.6.3 Participants and Experimental Set-up

The methodology closely followed that of Isola et al. [77] for measuring scene memorability. The experiment was set up as a game on Amazon Mechanical Turk, where workers were presented with a sequence of images, and had to press a key if they saw an image for the second time in the sequence. The repeated images were the target images of which we had 410, and the rest of the sequence was composed of "filler" images (i.e., the rest of all the single-panel visualizations in the database). Workers could complete up to 17 "levels" of the game, each comprised of about 120 images (targets and fillers) and taking about 4.8 minutes to complete. There was no change in difficulty across the levels, rather they were a way of partitioning the image stream and giving workers the option of taking a break (up to 5 minutes). Workers were paid $0.40 for each successfully-completed level, bringing their possible hourly wage up to almost $5. Workers could exit the game at any time, and were paid for the total amount of the game completed (including partially-completed levels). Upon completing a level, workers could see their average score (i.e., percent correctly-remembered images) for the level.

Target images were the ones for which we were interested in measuring memorability. The rest of the image sequence was filled with vigilance repeats and other filler images. Vigilance repeats consisted of an image repeated twice, with a spacing of 1-7 images, and were meant to be easy to detect. This was implemented to screen out workers that were not paying enough attention to the task. If a worker false-alarmed on more than 50% of the last 30 non-repeat images, the game would end, and the worker would be flagged. If flagged three times, a worker would be paid for the part of the game completed, and would be blocked from further participation.

Images in the sequence were presented for 1 second, with a 1.4 second gap between consecutive images. These sequences contained a different ordering of images for each worker. Images were presented a maximum of 2 times throughout the whole image sequence, and all the repeat images appeared 91-109 images apart. For our experiment, all images were resized to lie within a maximum dimension of 512×512 pixels (while preserving aspect ratios), so as to fit comfortably into a webpage containing the memorability game.

To begin the game, workers had to complete a practice trial with 30 images. Until a worker's miss rate on the practice fell below 50% and the false alarm rate fell below 30%, the worker could not continue on to the real game. A worker who failed the practice three times would be paid $0.02 for the practice, and blocked from the game.

On Amazon Mechanical Turk, we posted 276 HITs ("Human Intelligence Tasks"), each of which consisted of our game with 17 possible levels. To accept one of our HITs, a worker had to have an approval rate of over 95% in Amazon's system as a quality check. Of the workers who accepted our HITs, 261 passed the practice. Workers were able to accept the HIT multiple times and pick up where they left off (until either all 17 levels were completed or the worker failed the quality screening described above). Of the 261 workers, 57% saw more than 90% of the target images. The rest completed fewer of the levels. On average, we have 87 responses (SD: 4.3) per target image. The age range of our workers spanned 16 to 66 years, and the mean age was 33.1 (SD: 10.5). The race distribution was: 41.7% Caucasian, 37.5% South Asian, 4.2% African, 4.2% East Asian, 1.1% Hispanic, and 11.3% other/unreported. We did not collect any other demographic information from our workers.

Because we did not restrict participation in our task based on any worker demographics, we believe that we have sampled fairly from the Mechanical Turk worker population. Other studies have surveyed the Mechanical Turk population, and have determined the education and income of the workers to be quite diverse, reporting that the majority of workers earn roughly U.S. \$30*k* per year, and almost half have earned a bachelor's degree [115, 151].

5.6.4 Experimental Design & Analysis

Performance Metrics: Workers saw each target image at most 2 times (less than twice if they prematurely exited the game). We measure an image's *hit rate* (*HR*) as the proportion of times workers responded on the second (repeat) presentation of the image. In signal detection terms: $HR = \frac{HITS}{HITS+MISSES}$. We also measured how many times workers responded on the *first* presentation of the image. This corresponds to workers thinking they have seen the image before, even though they have not. This *false alarm rate* (*FAR*) is calculated: $FAR = \frac{FA}{FA+CR}$, where *FA* is the number of false alarms and *CR* is the number of correct rejections (the absense of a response).

For performing a relative sorting of our data instances we used the *d-prime* metric (otherwise called the *sensitivity index*). This is a common metric used in signal detection theory, which takes into account both the signal and noise of a data source, calculated as: $d' = Z(HR) - Z(FAR)$ (where Z is the inverse cumulative Gaussian distribution). A higher d' corresponds to a signal being more readily detected. Thus, we can use this as a *memorability score* for our visualizations. A high score will require the *HR* to be high and the *FAR* to be low. This will ensure that visualizations that are easily confused for others (high *FAR*) will have a lower memorability score.

Data Analysis: Of the 410 target visualizations selected for the memorability experiment, 17 were subsequently filtered out because their aspect ratios were deemed too skewed for the comparison to other visualizations to be fair. Visualizations with aspect ratio greater than 3:1 made the text hard to read, and pictographic elements hard to decipher. An initial analysis showed that these images ended up with high *FAR*, being confused for one another (losing their distinctiveness to aspect ratio similarities).

A set of analyses was run on the remaining images, whereby memorability score was plotted against various visualization attributes (Sec. 5.3). The plots were constructed by summarizing across all the visualizations, and also by individually considering visualization sources: government / world organizations, news media, infographics, and scientific publications. This was done to see whether there are differences in how the attributes correlate with memorability across different publication venues. To filter out the effect recognizable elements like people and objects have on memorability, the analyses were repeated by only considering visualizations that did not contain pictorial elements (filtered-out manually).

As d-prime is a normalized metric, corrected t-tests were applied in Sec. 5.7 to evaluate the statistical significance of the memorability scores of the different attributes, visualization types, and visualization sources.

5.7 Experimental Results & Discussion

5.7.1 Memorability comparisons

Our memorability experiment was designed to understand the memorability of visualizations within the context of other memorability studies, and treats the visualizations in our experiment as scenes (i.e., images or photographs). This memorability study design does not evaluate the cognitive impact or engagement of the visualization. Our experiment primarily presents baseline results to compare the memorability of visualizations to the memorability of scenes. In other words, we have measured how visualizations would be remembered if they were images.

We observed a mean *HR* of 55.36% ($SD = 16.51\%$) and mean *FAR* of 13.17% ($SD = 10.73\%$). For comparison, scene memorability has a mean *HR* of 67.5% ($SD = 13.6\%$) with mean *FAR* of 10.7% ($SD = 7.6\%$) [77], and face memorability has a mean *HR* of 53.6% ($SD = 14.3\%$) with mean *FAR* of 14.5%($SD = 9.9\%$) [9]. This possibly supports our first hypothesis that visualizations are less memorable than natural scenes. This demonstrates that there is memorability consistency with scenes, faces, and also visualizations, thus memorability is a generic principle with possibly similar generic, abstract, features.

We also measured the consistency of our memorability scores [9, 77]. By splitting the participants into two independent groups, we can measure how well the memorability scores of one group on all the target images compare to the scores of another group (Fig. 5.7). Averaging over 25 such random half-splits, we obtain Spearman's rank correlations of 0.83 for HR, 0.78 for FAR, and 0.81 for d-prime, the latter of which is plotted in Fig. 5.7. This high correlation demonstrates that the memorability of a visualization is a consistent measure across participants, and indicates real differences in memorability between visualizations. In other words, despite the noise introduced by worker variability and by showing different image sequences to different workers, we can nevertheless show that memorability is somehow intrinsic to the visualizations.

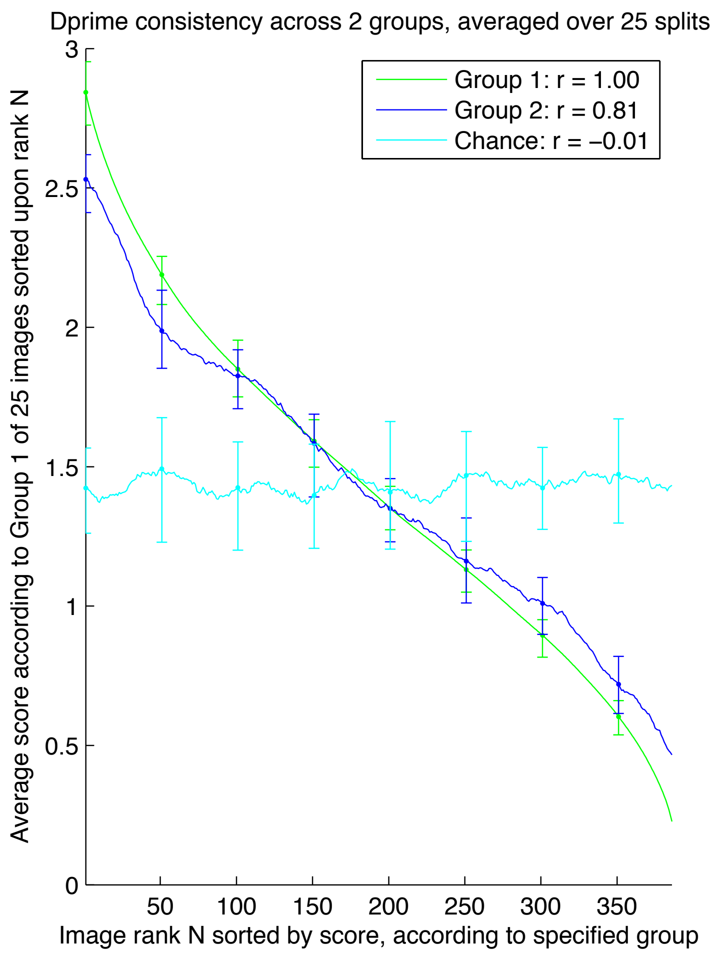

Figure 5.7: Participants were split into two independent sets, Group 1 and Group 2. Visualizations were ranked by memorability scores (d-prime) from participants in Group 1 (green line) or Group 2 (blue line) and plotted against the average memorability scores given by participants in Group 1. Plots are averaged across 25 such random splits. For clarity, we also convolved the resulting plots with a length-25 box filter along the x-axis. The cyan chance line was simulated by assigning the images random ranks (i.e., randomly permuting the x-axis). Error bars depict 80% confidence intervals. Note that the scores of the two participant halves are highly correlated over the 25 random half-splits.

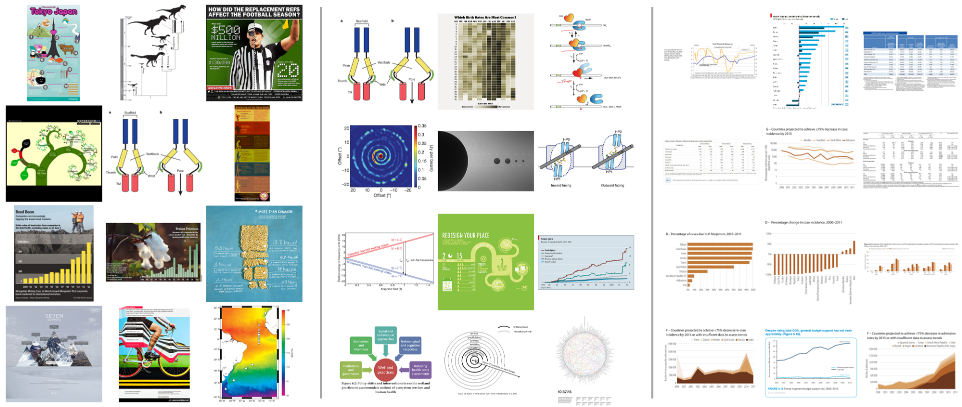

Figure 5.8: Left: The top twelve overall most memorable visualizations from our experiment (most to least memorable from top left to bottom right). **Middle:** The top twelve most memorable visualizations from our experiment when visualizations containing human recognizable cartoons or images are removed (most to least memorable from top left to bottom right). **Right:** The twelve least memorable visualizations from our experiment (most to least memorable from top left to bottom right).

5.7.2 Visualization Attributes

Of our 410 target visualizations, 145 contained either photographs, cartoons, or other pictograms of human recognizable objects (from here on out referred to broadly as "pictograms"). Visualizations containing pictograms have on average a higher memorability score (Mean (M)=1.93) than visualizations without pictograms ($M = 1.14, t(297) = 13.67, p < 0.001$). This supports our second hypothesis. Thus not all chart junk is created equal: annotations and representations containing pictograms are across the board more memorable. However, this is not too surprising as we are evolved to see, segment, and recognize natural objects. Thus an image, or image of a visualization, containing a human recognizable object will be easily recognizable and probably memorable.

Due to this strong main effect of pictograms, we examined our results for both

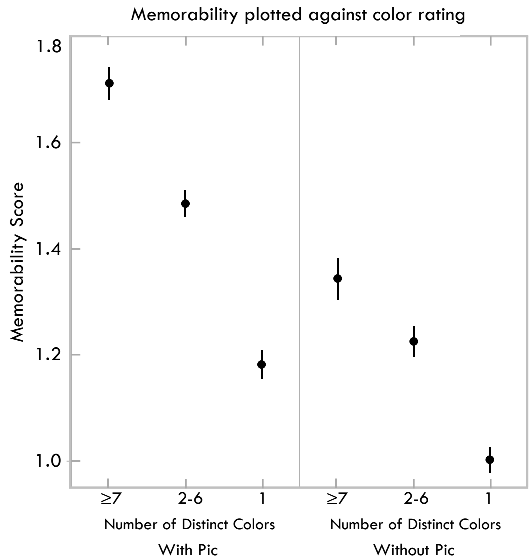

Figure 5.9: Memorability scores for visualizations based on the number of colors it contains. On the left is all visualizations, and on the right visualizations with pictograms removed.

the cases of visualizations with and without pictograms. As shown in the left-most panel of Fig. 5.8, all but one of the overall top most memorable images (as ranked by their d-prime scores) contain human recognizable pictograms. The one visualization without a human recognizable image, the molecular diagram in the middle of the second row, is the most memorable image of our non-pictogram visualizations (see Fig. 5.8, middle panel). The least memorable visualizations are presented in the right-most panel of Fig. 5.8.

As shown in Fig. 5.9, there is an observable trend of more colorful visualizations having a higher memorability score: visualizations with 7 or more colors have a higher memorability score ($M = 1.71$) than visualizations with 2-6 colors ($M = 1.48, t(285) = 3.97, p < 0.001$), and even more than visualizations with 1 color or black-and-white gradient ($M = 1.18, t(220) = 6.38, p < 0.001$). When we remove

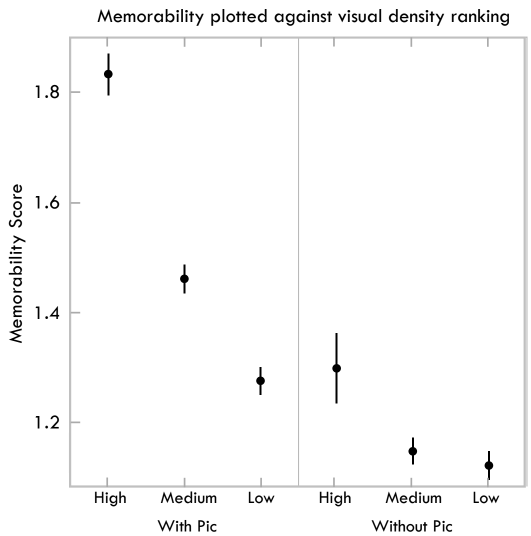

Figure 5.10: Memorability scores for visualizations based on visual density. On the left is all visualizations, and on the right visualizations with pictograms removed.

visualizations with pictograms, the difference between visualizations with 7 or more colors ($M = 1.34$) and those that have only 1 color ($M = 1.00$) remains statistically significant ($t(71) = 3.61, p < 0.001$).

Considering all the visualizations together, we observed a statistically significant effect of visual density on memorability scores with a high visual density rating of "3" ($M = 1.83$), i.e., very dense, being greater than a low visual density rating of "1" ($M = 1.28, t(115) = 6.08, p < 0.001$) as shown in Fig. 5.10.

We also observed a statistically significant effect of the data-to-ink ratio attribute on memorability scores with a "bad" ($M = 1.81$), i.e., low data-to-ink ratio, being higher than a "good" rating ($M = 1.23, t(208) = 6.92, p < 0.001$) as shown in Fig. 5.11. Note that using a corrected t-test, we also arrive at the results that the 3 levels of data-ink ratio are pairwise significantly different from each-other.

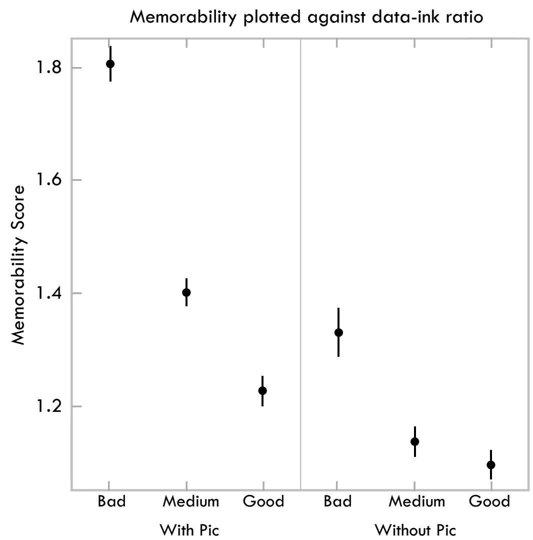

Figure 5.11: Memorability scores for visualizations based on the data-to-ink attribute ratings. On the left is all visualizations, and on the right visualizations with pictograms removed.

Summarizing all of these attribute results: higher memorability scores were correlated with visualizations containing pictograms, more color, low data-to-ink ratios, and high visual densities. This supports our third hypothesis, and refutes our fourth and fifth hypotheses. However, as discussed in Sec. 5.7.1, we tested the memorability of visualizations as images and not the comprehension of the visualizations. Thus, looking at these visualizations as *images* and not data encodings, these attributes increased their memorability.

As shown in Fig. 5.12, diagrams were statistically more memorable than points, bars, lines, and tables. These trends remain observable even when visualizations with pictograms are removed from the data. Other than some minor ranking differences and addition of the map category, the main difference is in the ranking of the table visualization type, which without pictograms becomes least memorable.

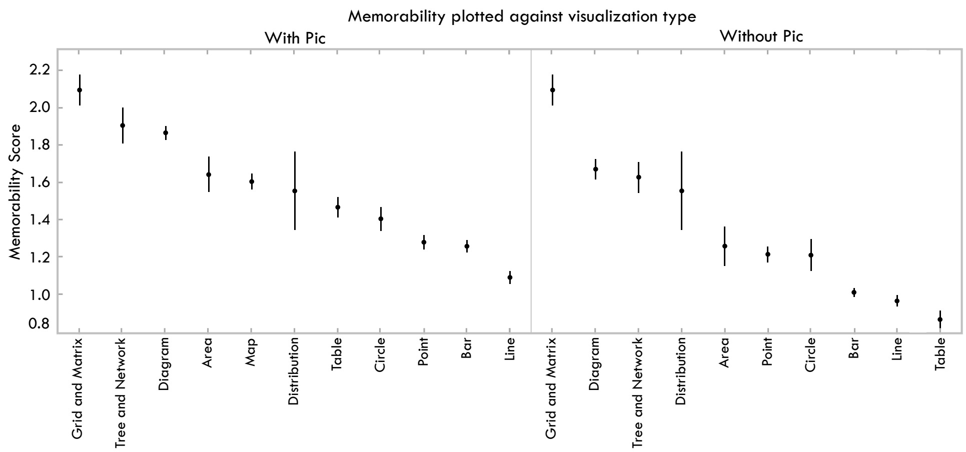

Figure 5.12: Memorability scores for visualizations based on visualization type. On the left is all visualizations, and on the right visualizations with pictograms removed.

The middle panel of Fig. 5.8 displays the most memorable visualizations that do not contain pictograms. Why are these visualizations more memorable than the ones in the right-most panel? To start with, qualitatively viewing the most memorable visualizations, most are high contrast. These images also all have more color, a trend quantitatively demonstrated in Sec. 5.7.2 to be correlated with higher memorability. As compared to the more subdued less memorable visualizations, the more memorable visualizations are easier to see and discriminate as images. Another possible explanation is that "unique" types of visualizations, such as diagrams, are more memorable than "common" types of visualizations, such as bar charts. This trend is also evident in Fig. 5.12 in which grid/matrix, trees and networks, and diagrams have the highest memorability scores. This observation refutes our sixth hypothesis. Examples of these unique types of visualizations are each individual and unique, whereas bar charts and line graphs are uniform with limited variability in their visual encoding methodology. Previously it has been shown that an item is more likely to interfere with another item if it has similar category or subordinate category information, but unique exemplars of objects can be encoded in memory quite well [88]. This supports our findings that show high FAR and low HR for table and bar visualizations, which both have very similar visuals within their category (i.e., all the bar charts look alike). Another contributing factor is that target visualizations represented a distribution of types found "in the wild." Thus, of the 410 target visualizations, trees and networks totaled 11 targets and grid/matrix totaled 6 targets. Thus their low frequency may have contributed to their uniqueness. However, this was not the case for diagrams, which constituted 57 of the target visualizations.

Another possible explanation is that visualizations like bar and line graphs are just not natural. If image memorability is correlated with the ability to recognize natural, or natural looking, objects then people may see diagrams, radial plots, or heat maps as looking more "natural". Previous work has shown that people can rapidly identify common objects or concepts, referred to as processing fluency, and that prior expectations will influence a person's performance [138]. Since we are mostly attuned to natural scenes, it makes sense that some of the top memorable visualizations look closer to "nature" than the others. In these terms, we see that people may have more perceptual fluency with visualizations that at first glance appear to be more "natural" and that this fluency may be influencing memorability.

One common visual aspect of the most memorable visualizations is the prevalence of circles and round edges. Previous work has demonstrated that people's emotions are more positive toward rounded corners than sharp corners [10]. This could possibly support both the trend of circular features in the memorable images as well as the concept of natural-looking visualizations being more memo-

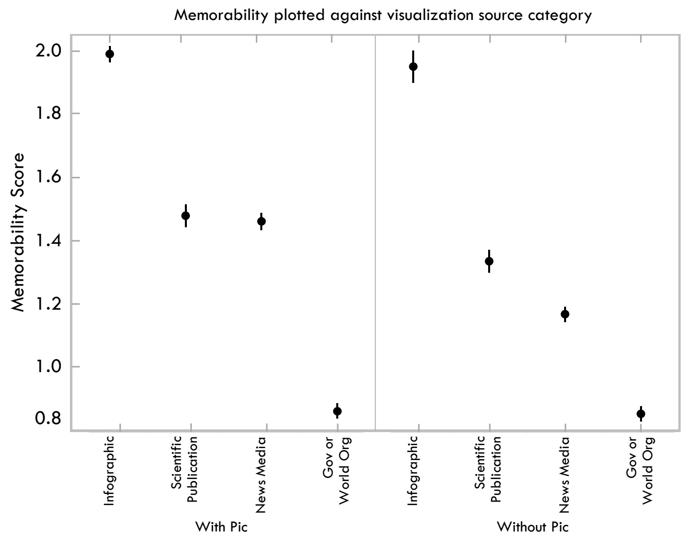

Figure 5.13: Memorability scores for visualizations based on original source category. On the left is all visualizations, and on the right visualizations with pictograms removed.

rable since "natural" things tend to be round.

5.7.3 Visualization Sources

As shown in Fig. 5.13, regardless of whether the visualizations did or did not include pictograms, the visualization source with the highest memorability score was the infographic category ($M = 1.99, t(147) = 5.96, p < 0.001$ when compared to the next highest category, scientific publications with $M = 1.48$), and the visualization source with the lowest memorability score was the government and world organizations category ($M = 0.86, t(220) = 8.46, p < 0.001$ when compared to the next lowest category, news media with $M = 1.46$). These results were significant according to corrected t-tests. Note that these statistically-significant trends hold even with visualizations containing pictograms removed. In fact, with

Figure 5.14: The top ten most memorable visualizations for each of the four visualization source categories: infographic (top left), scientific publications (top right), news media (bottom left), and government / world organization (bottom right). In each quadrant, the visualizations are ordered most to least memorable from top left to bottom right.

pictograms removed, scientific publications ($M = 1.95$) become significantly more memorable than news media ($M = 1.17, t(23) = 6.92, p < 0.001$). The top ten most memorable visualizations from each source category are shown in Fig. 5.14.

A few things to bear in mind: first of all, as previously stated, this memorability study examined the memorability of visualizations as if they were images and not memorability based on engagement and comprehension of the visualization. Secondly, the visualizations in this category were drawn from Visual.ly with a more design focused venue and audience where the visualizations are intentionally created to be flashy and include stylized elements. Another factor is that visualizations are submitted to Visual.ly and are pre-judged by people before being published. In contrast, the other venues contain a more unbiased and not pre-judged selection of visualizations. The visualizations for a source such as Visual.ly, or even other news media sites, are competing for a viewer's attention. Thus they will probably be more likely, intentionally or unintentionally, to use bright, bold, pictorial visual elements to grab a reader's attention. Thus this type of publication venue's motivational bias may translate into design features that lead to higher memorability.

Another possible influence of visualization source on memorability score is publication venue specific aesthetics. Many visualizations, particularly those from the news media and government sources, tend to publish with the same visual aesthetic style. This may be due to either the venue maintaining a consistent look so viewers will automatically recognize that a visualization was published by them, or because they have editorial standards to create visualizations that appear similar. This may have a negative impact on memorability scores because visualizations of similar aesthetics lack uniqueness. This may be a contributing factor to the observed trend (see Fig. 5.13) that visualization sources that have non-uniform aesthetics tend to have higher memorability scores than sources with uniform aesthetics. This observation refutes our last hypothesis that visualizations from scientific publications are less memorable. This may also be due to the fact that visualizations in scientific publications have a high percentage of diagrams (Fig. 5.6), similar to the infographic category.

5.8 Conclusions & Future Work

The results of our memorability experiment show that visualizations are intrinsically memorable with consistency across people. They are less memorable than natural scenes, but similar to images of faces, which may hint at generic, abstract, features of human memory. Not surprisingly, attributes such as color and the inclusion of a human recognizable object enhance memorability. And similar to previous studies we found that visualizations with low data-to-ink ratios and high visual densities (i.e., more chart junk and "clutter") were more memorable than minimal, "clean" visualizations. It appears that we are best at remembering "natural" looking visualizations, as they are similar to scenes, objects, and people, and that pictorial and rounded features help memorability.

More surprisingly, we found that unique visualization types (pictoral, grid/matrix, trees and networks, and diagrams) had significantly higher memorability scores than common graphs (circles, area, points, bars, and lines). It appears that novel and unexpected visualizations can be better remembered than the visualizations with limited variability that we are exposed to since elementary school. In hindsight this finding is consistent with results for natural scenes and objects.

Our results seem to validate the opinions of proponents on both sides of the chart junk debate. Edward Tufte says: "All the history of information displays and statistical graphics – indeed of any communication device – is entirely a progress of methods for enhancing density, complexity, dimensionality, and even sometimes beauty." [174] And Nigel Holmes states: "As long as the artist understands that the primary function is to convey statistics and respects that duty, then you can have fun (or be serious) with the image; that is, the form in which these statistics appear." [70] We believe that visualizations are what Alberto Cairo calls a Functional Art: "something that achieves beauty not through the subjective, freely wandering self-expression of the painter or sculptor, but through the careful and restrained tinkering of the engineer." [32] But it appears that the artist and designer can have a big influence in making visualizations more memorable.

Understanding what makes a visualization memorable is only the first step to understanding how to create effective data presentations. Making a visualization more memorable means making some part of the visualization "stick" in the viewers mind. We do not want just any part of the visualization to stick (e.g., chart junk), but rather we want the most important relevant aspects of the data or trend the author is trying to convey to stick. If we can accomplish this, then we will have a method for making data more memorable. This will have diverse applications in education, business, and more generally, in how data is presented to wide audiences.

In future work we hope to gain further understanding of the memorability of visualizations. This would include expanding our visualization database in order to gain an even more diverse real world sample, annotating more of the images with all visualization types and attributes of our taxonomy in order to better understand the memorability subtleties of specific types or subtypes of visualizations, annotating visualizations with more fine-grained definitions and measures of visual density, and investigating how memorability is impacted by multiple visualizations (e.g., small multiples or multi-panel visualizations). We plan to investigate the effect of time on memorability of visualizations, and investigate whether certain visual features stick in the viewers mind longer than others. A particular category worth investigating further is pictograms. We would like to break this category down into subtypes to look for specific effects on memorability. We also hope to show in future work that memorability – i.e., treating visualizations as scenes – does not necessarily translate to an understanding of the visualizations themselves. Nor does excessive visual clutter aid comprehension of the actual information in the visualization (and may instead interfere with it). Finally, we hope to conduct eye movement studies to identify the parts of visualizations used for memory or comprehension.

Having a more solid understanding of the memorability of visualizations will also allow us to carefully craft future studies to ask the more important and interesting questions of what makes a visualization comprehensible, engaging, or impactful. With a more solid grasp of what visual elements impact memorability at a low level, we can control for them at a higher level so as not to interfere with other factors in future experiments. We will then be able to start answering the larger questions of how to design effective visualizations.

6

Eye-tracking Study for Visualization Recognition and Recall

In this chapter we present a follow-on evaluation study to that described in the previous chapter. In Chapter 5 we established that visualizations are inherently memorable and which visual features or visualization types contribute to increased memorability. In this chapter we go deeper and further into this theory and conduct an eye-tracking evaluation in order to understand exactly which elements of a visualization contribute to memorability as well as a visualization's recognizability and recall.

6.1 Motivation

Understanding the basic cognitive and perceptual properties of a visualization is essential for effective data presentation as well as communication to the viewer. Memorability, a basic cognitive concept, has important implications for both the design of visualizations that will be remembered but also lays the groundwork for understanding higher cognitive functions such as comprehension. In our previous study, as described in Chapter 5, the memorability scores for hundreds of real-world visualizations were collected on Amazon's Mechanical Turk (AMT). The results of this research demonstrate that visualizations have inherent memorability and are reproducible across different groups of observers. We also found that the most memorable visualization types are those which are visually distinct (e.g., diagrams, tree and network diagrams, etc.), and that elements such as color, visual complexity, and recognizable objects increase a visualization's memorability.

Here we move beyond memorability and study which *features* of a visualization contribute to memory, recognition, and recall. What information is stored and recalled when viewing a visualization for a few seconds? In the previous study, each visualization was presented for 1 second. However, when looking at an image for a few seconds, working memory comes into play which enables the integration, processing, disposal, and retrieval of information. Thus in this study we explore how people encode, recognize, and recall visualizations. As markers of encoding and recognition we recorded eye movements of observers and looked at the patterns of gaze when observers study the visualization for the first time (*encoding*

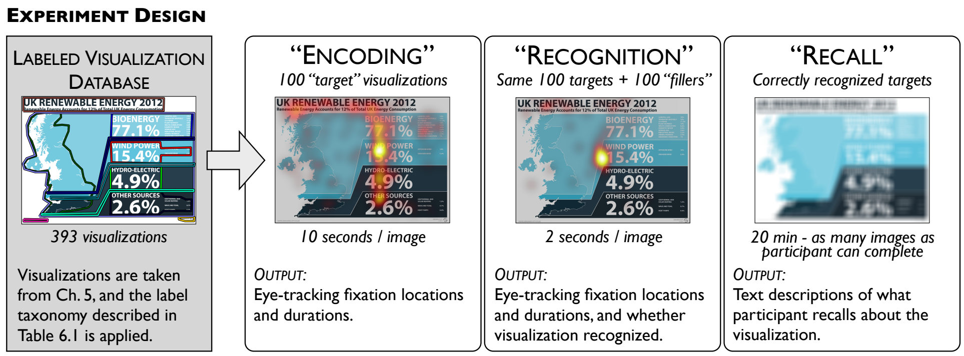

Figure 6.1: Illustrative diagram of the experiment design. From left to right: the elements of the visualizations are labeled and categorized, eye-tracking fixations are gathered for 10 seconds of "encoding", eye-tracking fixations are gathered while visualization recognizability is measured, and finally participants provide text descriptions of the visualizations based on blurred representations to gauge recall.

phase) and indicate whether they recognize the image or not during a *recognition phase* (see Fig. 6.1). In the third *recall phase*, observers were asked to describe what they remember about a specific visualization (Fig. 6.1).

In this chapter we present the results of an experiment that measures visualization recognizability as well as evaluates what people remember about a visualization. We executed an eye-tracking lab study with a labeled database of hundreds of real-world visualizations to determine which elements of a visualization draw attention. We also investigated which visualization elements people attend to in order to recognize and recall a visualization through the analysis of recognition fixations, and text descriptions composed by study participants.

This work represents the first eye-tracking study to investigate which types of visual components contribute to visualization recognition and recall. In addition, we present an analysis of the labeled visualizations in order to characterize visualization design characteristics across different venues. Finally, based on the results of our experiment, we present a list of conclusions as conceptual tools for visualization design.

6.2 Related Work

Memorability of Visualizations: As described in Sec. 6.1 and Ch. 5, we evaluated the memorability of visualizations using thousands of un-edited visualizations. We found that some visualization types are more memorable than others, and particular visual elements (e.g., human recognizable objects, color, etc.) seem to increase a visualization's memorability. In this study we build on this previous research, as well as the related works discussed in Sec. 2.1, and move beyond basic memorability by studying what visualization types and features contribute to memory recognition (i.e., when a person remembers seeing a particular visualization) and recall (i.e.,when a person describes what he or she remembers from the visualization).

Eye-tracking Evaluations in Visualization: Eye-tracking evaluations can be an effective tool for understanding how a person views and visually explores a visualization. It has been used in the visualization community for evaluating specific visualization types such as graphs [71, 72, 90, 134], tree diagrams [30], and parallel coordinates [161], the comparison of multiple types of visualizations [54], and the evaluation of visualization tool interactions [85]. There has also been research in the area of understanding different types of tasks and visual search strategies for visualizations through the analysis of eye-tracking fixation patterns as well as insights into cognitive processes [134, 137]. The work presented in this chapter does not focus on specific tasks, nor a specific type of visualization, but rather utilizes eye-tracking for fixation location and duration analysis on hundreds of labeled and categorized visualizations with dozens of study participants. Through this analysis within the context of our experimental design, we are able to more deeply understand the specific cognitive processes of recognition and recall of visualizations.

6.3 Study Overview

In following sections we describe the detailed labeling of visualizations for our experiment followed by a discussion of the three parts of the experiment and results. Following along the experiment design workflow illustration in Fig. 6.1, the first step was to manually label each visual element in the 393 visualizations used as *target* visualizations in the previous memorability study (Chapter 5). These labeled visualizations were then used in our experiment which was composed of three parts: the *encoding* phase, the *recognition* phase, and the *recall* phase. In the encoding phase, participants viewed 100 target visualizations for 10 seconds each in an eye-tracking set-up. This enabled the collection of glance fixations in order to examine which elements a person focuses on when visually exploring a visualization.

The next portion of the experiment, the recognition phase, is analogous to the memorability experiment in Chapter 5 except that the target visualizations have been previously viewed for 10 seconds, instead of 1 second, thus engaging working memory. Eye-tracking data was collected as participants viewed the labeled target visualizations, mixed-in with *filler* previously unseen visualizations, for 2 seconds each. The participants clicked the space-bar key in order to identify whether they had previously seen the visualization.

Finally, in the *recall* phase, participants were presented with the target visualizations they correctly identified during the recognition phase in a randomized order and asked to “Describe the visualization in as much detail as possible.” The visualizations presented during this phase were blurred so that they were still recognizable, but no new information obtainable as the text was unreadable. This last experiment phase does not explicitly evaluate a participant’s comprehension of a visualization, but does provide insight into what visualization elements, types, and concepts persisted in their working memory.

6.4 Data Collection and Annotation

For this work we utilized the database of visualizations gathered for our previous memorability study in Chapter 5. The database was generated by scraping multiple sources of real-world visualization publication venues online covering government reports, infographic blogs, news media, and scientific journals. The diversity and distribution of these visualizations represent a broad look of data visualization “in the wild.” For our study, the same 393 target visualizations from the previously published study were used along with 393 visualizations selected from the remaining *single* visualizations in the collection as filler images for our experiment (see Sec. 6.6.3).

In order to gain deeper insight into precisely which elements of a visualization may effect its memorability, recognition, and recall, we manually labeled each of the visual elements in the 393 target visualizations. The labels were applied using the LabelMe system [152]. The labels were hand-drawn by three students who had all completed the Harvard University introductory course in visualization. The labels were reviewed for accuracy and consistency, and corrected as needed by a visualization expert. As part of the LabelMe system, the labels were recorded as

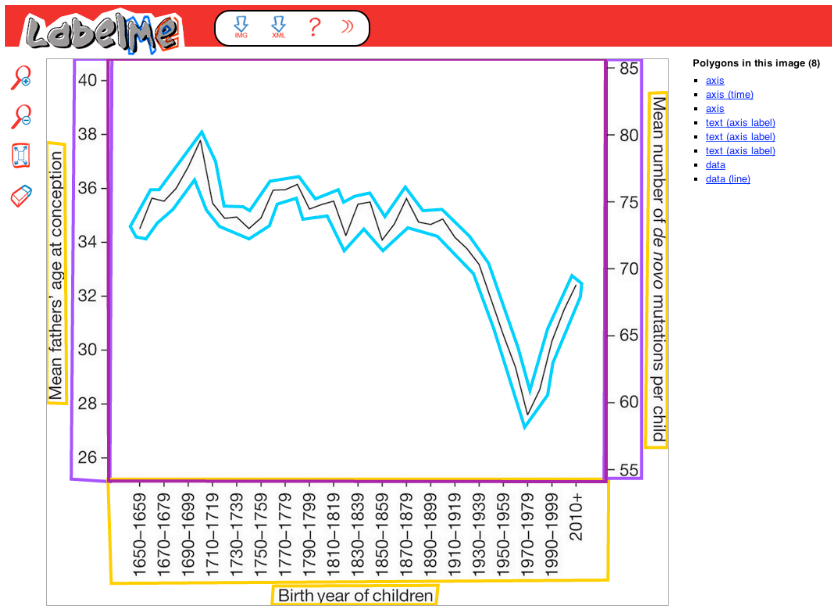

Figure 6.2: Example labeled visualization in the LabelMe system [152]. This line graph from a scientific journal has 8 labeled elements including where the data, axes, and title are located.

polygons in xml and then converted to binary masks for further analysis. Examples of the labeled visualizations are shown in the leftmost panel of Fig. 6.1 and in Fig. 6.2.

The labeling taxonomy was based on the visualization taxonomy created in Chapter 5. As described in Table 6.1, the labels classify the elements in the visualization to be either encoding data, data related components (e.g., axes, annotations, legends, etc.), textual elements (e.g., title, axis labels, paragraphs, etc.), human recognizable objects, or graphical elements with no data encoding function. The labels were allowed to overlap each other in the visualization (e.g., text label or annotation on top of a graph).

Table 6.1: Summary of our visualization labeling taxonomy for the eye-tracking experiment. The data subtypes taxonomy is taken from Chapter 5.

LABEL	[OPTIONAL SUBTYPES] DESCRIPTION
Annotation	[Arrow] Outline of any visual elements annotating the data. A specific subtype of "arrow" was included to denote whether the annotation was in the form of an arrow.
Axis	[Time] Outline of where an axis is located including any tic marks and numeric values along the axis. A specific subtype of "time" was included to denote an axis involving time.
Data	Outline of where the full data plot area is located (e.g., the area between the x-axis and y-axis in a 2D plot).
Data (type)	[Area, Bar, Circle, Diagram, Distribution, Grid & Matrix, Line, Map, Point, Table, Text, Trees & Networks] Outline of where the actual data values are visually encoded (e.g., bars in a bar graph, points in a scatterplot, etc.).
Graphical Element	Outline of any visual elements that are not related to the visual representation or description of the data.
Legend	Outline of any legends or keys that explain the data's visual encoding (e.g., color scales, symbol representations, map legends, etc.).
Object	[Photograph, Pictogram] Outline of any human recognizable objects in the image. Objects are either realistic in representation (photograph) or abstract drawings (pictogram). Descriptions of each object were also recorded.
Text	[Axis Label, Header Row, Label, Paragraph, Source, Title] Outline of any text in the image. Subtypes cover all the common representations from prose to labels.

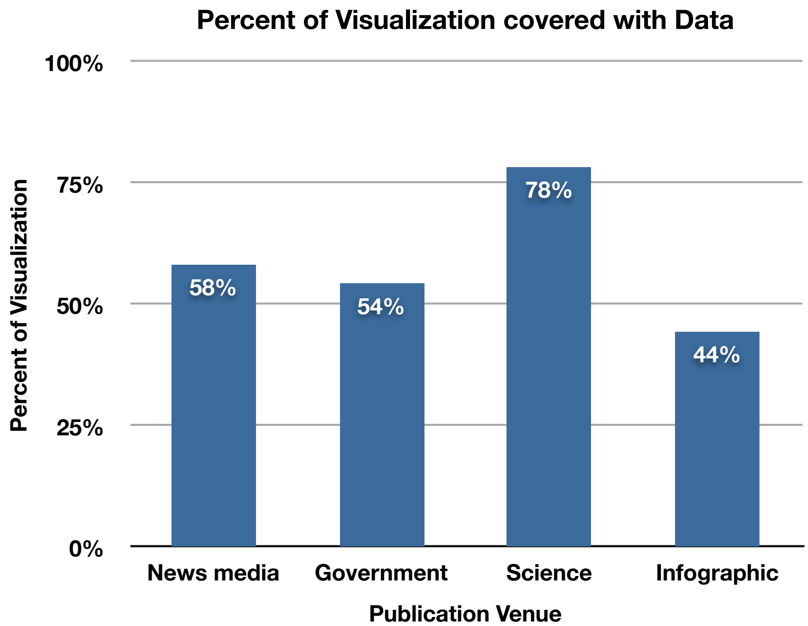

Figure 6.3: Percentage of visualization pixel area covered by the *data* label.

6.5 Analysis of Labeled Visualizations

The labeled visualizations enable us to study the distribution and type of visual elements employed across publication venues as well as visualization types. Examining the percentage of the image area covered by the *data* label (i.e., the image area devoted to displaying data), as shown in Fig. 6.3, we see that the proportion of the image area devoted to data display is highest for the scientific journal visualizations. This is probably due to the specific publishing context of scientific journal figures in which the visualization is part of a paper narrative thus requiering less accompanying text and other elements in the figure. Another contextual factor to consider is the page limitations and high prices of journal paper publishing for the author, thus the motivation to maximize data display per visualization.

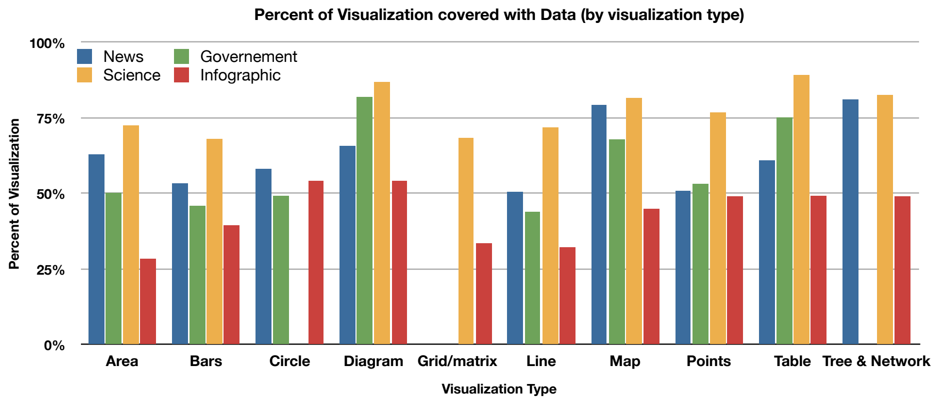

Figure 6.4: The percent of image area covered by the *data* label for each visualization type per visualization source. The infographic visualizations in general have the smallest image area devoted to displaying data, whereas the visualization from scientific journals have the highest. The visualization types with the largest areas for data display are diagrams, maps, and tables.

If we break down this measure of image area for data display by visualization type, as shown in Fig. 6.4, we see that diagrams, maps, and tables cover a larger percentage of the image area than other visualization types. These types of visualizations tend to be more "all inclusive" with annotations and text labels more commonly incorporated into the data representation itself thus requiring less of the image area around the data plot for additional explanations. Worth noting is that the infographic visualizations for most visualization types have a smaller image area for data display as compared to the other publication venues.

Another interesting observation is the distinct difference in the average total number of labels in a visualization between the different sources. Visualizations from government sources have on average 11.98 labeled elements per visualization, significantly fewer on average as compared to the other visualization sources ($t(152) = 5.76, p < 0.0001$). In contrast, visualizations from infographic sources have nearly twice as many elements ($M = 38.66$) as compared news media

($M = 19.61$) ($t(211) = 5.72, p < 0.0001$) and scientific visualizations ($M = 18.57$). The additional elements in the infographic visualizations are mostly in the form of more text elements, objects, and graphical elements around the data.

Finally, there is a distinct difference between publication venues in regards to the percentage of the visualization's area covered in human recognizable objects. There are no such objects in the government published visualizations, and the percentages are less for scientific journal visualizations ($M = 14\%$) as compared to news media ($M = 25\%$) and infographic ($M = 20\%$) visualizations. As shown in the word cloud visualization in Fig. 6.5, the human recognizable objects are primarily in the form of company logos. Another large source of pictogram objects are international flags commonly used in the news media visualizations to denote different countries. Other common visual elements include pictograms or photographs of human representations and computer/technology depictions.

6.6 Eye-tracking Experiment Overview

6.6.1 Experiment Set-up & Participants

As discussed in Sec. 6.4, for the 393 target visualizations which we have memorability scores (Chapter 5) we labeled each of the visual elements in the visualizations. We also carefully selected 393 visualizations that match the exact distribution of visualization type and original source as the target visualizations from the database of single visualizations in Chapter 5 to use as filler visualizations during the recognition portion of the experiment. All of the target and filler visualizations were resized, while preserving aspect ratios, so that their maximum dimension was 1000 pixels.

Figure 6.5: Word cloud of the individual words in the descriptions of each human recognizable object in the labeled database for the target visualizations. The size of the word is proportional to its frequency in the database.

For the eye-tracking portions of our experiment, we used an SR Research Eye-Link1000 desktop eye-tracking system equipped with a chin-rest mount 22 inches from a 19 inch CRT monitor with a resolution of 1280 x 1024 pixels. At the beginning of each experiment session, participants performed a randomized 9-point calibration and validation procedure. Each time a participant took an optional break during the encoding and recognition phases of the experiment a drift check was performed and, if necessary, a recalibration was performed.

A total of 25 participants (14 females, 11 males) participated in the experiment. All of the participants were students affiliated with MIT with an age range of 18-32 ($M = 22.9, SD = 4.2$). All participants had normal or corrected-to-normal vision, and normal color vision. Each experiment session covered $\sim$25% of the target visualizations, thus each individual could participate in the experiment up to four times. On average each participant completed 1.92 experiment sessions with 6 participants completing all 4 experiment sessions. A single session of the experiment lasted $\sim$1 hour. Each participant was monetarily compensated for their time.

6.6.2 Encoding Phase

The first portion of the experiment was the encoding phase which lasted 20 minutes. As illustrated in Fig. 6.1, the participants were each shown 100 visualizations that were randomly selected from the labeled 393 target visualizations. For this phase of the experiment, participants were set-up and seated in the eye-tracking system and shown each visualization on the screen for 10 seconds. Each consecutive visualization was separated by a 0.5 second fixation cross to clear their field of view. Every 10 visualizations, the subject was given an opportunity to take a break (e.g., to stretch, rest their eyes, etc.). The visualization display duration of 10 seconds was chosen based on a pilot study in which different encoding durations where explored. A 10 second duration proved to be of sufficient length for a participant to read the visualization's title, axes, annotations, etc. as well as explore the data encoding, and was short enough to avoid too much redundancy in re-fixations as well as explorative strategies.

The measures collected during this phase were the eye-tracking fixation locations and durations.

6.6.3 Recognition Phase

The second portion of the experiment was the recognition phase which lasted 10 minutes. The participants were each shown the same 100 visualizations from the encoding phase as well as 100 filler visualizations. These 200 visualizations were presented to the participant in a random permutation for 2 seconds each with a 0.5 second fixation cross between consecutive visualizations. Participants were instructed to press the space bar anytime they recognized a visualization from the previous experimental phase. The participant could press the space bar as long as the visualization was still on the screen. The participants would receive feedback in the form of a message right before the fixation cross screen if they were correct or incorrect. Participants could take an optional break every 20 visualizations.

The measures collected during this phase were the eye-tracking fixation locations and durations, the number of correctly recognized visualizations (*HITs*), and how many seconds it took for a participant to click the space bar.

6.6.4 Recall Phase

The third and final portion of the experiment was the recall phase which lasted 20 minutes. Participants' gazes were not recorded with the eye-tracking system thus they could sit normally. In this phase, all of the visualizations that the participant correctly recognized in the previous phase of the experiment were presented in a randomized sequence on the screen. Each visualization was blurred by a Gaussian with a width of 40 pixels (which was found to be appropriate through a pilot study) in order to make the text unreadable. The purpose of the blurred visualizations was to present the participant with a visualization that was recognizable as being one of the correctly recognized visualizations from the previous experiment phase, but for the visualization to not have enough visual detail to enable the extraction of new information. Each visualization was presented at 60% of its original size.

Next to each blurred visualization an empty text box was presented on the screen with the following instructions: "Describe the visualization in as much detail as possible." The goal of this instruction was to illicit from the participant as much information about the visualization as they could recall from memory. This question was not explicitly designed to gauge comprehension of the visualization (e.g., what was the main point of the visualization, what are the main trends, etc.), but rather cognitive recall. Participants were given 20 minutes to write as many descriptions as possible. There was no limit to how much time or text length was spent on each visualization, nor any requirement to complete a certain number of visualizations. Each participant worked at his or her own speed and level of detail. Participants were also allowed to skip visualizations for which they could not construct a description.

The measures collected during this phase were the participant generated text descriptions of what they could recall of a given visualization, the time to complete each text description, and whether a visualization was skipped.

6.6.5 Performance Metrics

We compute multiple fixation measures by intersecting fixation locations on a visualization with the labeled visual elements (Table 6.1). This allows us to determine when fixations land within a given element. Note that a single fixation can land on several elements at once (e.g., an annotation on a graph). In this case, we count the fixation as belonging to all of those elements. We collect all of a viewer's fixations during a particular viewing period (i.e., 10 seconds for encoding, 2 seconds for recognition), and we discard as noise fixations lasting less than 150 milliseconds.

Below we summarize the primary metrics calculated in order to evaluate the eye-tracking data:

Total fixation time (TFT): The total duration of all of a viewer's fixations that land on a given visual element throughout the entire viewing period.

Fixation time per unit area (FTA): A viewer's total fixation time (TFT) divided by the area (in pixels) of a visual element.

Diversity of fixations (DOF): The number of unique elements fixated upon by a viewer during the entire viewing period.

Inter-element fixations (IEF): The number of times a viewer fixates on a different set of visual elements from one fixation to the next. Some of the elements fixated can be the same, as long as the whole set is different.

Re-fixations: The number of times a viewer returns to an element during the entire viewing period (including the first time the element is fixated). Consecutive fixations on the same element are not counted.

(Note that for both TFT and FTA, if we were interested in multiple elements at once (e.g., all text elements: title, label, paragraph, etc.), then we would compute the TFT and FTA per element and average over all elements to produce a single number for a particular viewer and a particular visualization.)

The above measures are averaged across viewers and different sets of visualizations, depending on the analysis. We computed Bonferonni-corrected t-tests for all our measurements.

We also compute the recognition hit rate (*HR*) for each visualization. This is the fraction of participants who correctly recognized a visualization when they saw it during the recognition phase of the experiment. This value ranges from 0 to 1. Note that in the memorability study presented in [4], *dprime* was used as a measure of memorability to take into account the false alarms during encoding. In this previous study, encoding and recognition trials were intermixed and treated the same (1 second for each presentation of the visualization). In the new experiment discussed in this paper the false alarms during encoding are not defined so, instead of using the dprime measure, we use the pure HIT rate (HR), the proportion of individuals who remember the images during the recognition trials, as our measure of recognizability.

6.7 Experimental Results and Discussion

6.7.1 Eye-tracking Fixation Metrics for Encoding

TFT: Across all visualization types and publication sources in our study, we observed the total fixation time to be highest for the data label ($M = 8752.55ms$, $t(8904) =$

$62.97, p < 0.0001$), followed by text ($M = 3284.88ms, t(5834) = 14.75, p < 0.0001$), and then human recognizable objects ($M = 2269.92ms, t(3844) = 11.59, p < 0.0001$). Thus the participant's viewing time was spent primarily on the data, followed by reading text, and then viewing human recognizable objects. Although objects are the third most viewed label across all visualizations, it should be noted that labels could overlap and a significant portion of the human recognizable objects' function is to encode the actual data (e.g., diagrams, pictorial unit charts, etc.). Examining the fixation duration by each of the visualization venues, all of the source categories have this same trend of data with the highest fixation duration followed by the text.

FTA: The fixation time per unit area across all visualizations is highest for text ($M = 0.36, t(5834) = 12.09, p < 0.0001$), human recognizable objects ($M = 0.18, t(3103) = 3.62, p < 0.0003$), and annotation labels ($M = 0.14, t(4183) = 8.45, p < 0.0001$). Thus, when normalizing for size of the element, text is the visual element people spend the most time viewing in a visualization during encoding.

DOF: We observed the highest diversity of fixations with visualizations from infographic sources which have an average DOF of 11.37 elements ($t(1947) = 10.86, p < 0.0001$). The source with the lowest average DOF is government sources ($M = 6.29, t(2210) = 20.77, p < 0.0001$). This indicates that participants viewed on average more elements in the infographics as compared to the other visualization sources. This is probably due to infographic visualizations having nearly twice as many visual elements on average (M = 38.66) as compared to news media ($M = 19.61$) ($t(211) = 5.72, p < 0.0001$) or scientific ($M = 18.57$) visualizations. In contrast, visualizations from government sources have on average nearly half ($M = 11.98$) the visual elements as compared to scientific visualizations ($t(152) = 5.76, p < 0.0001$). Thus the more visual elements in a visualization, the greater number of elements a person will view. However, given these results, a participant on average only viewed 29% of the elements in an infographic as compared to 53% of the elements in a government visualization. Thus a participant was able to more fully explore the elements in the non-infographic visualizations.

IEF: The visualizations from government publications have the most inter-element fixations ($M = 5.10, t(2210) = 5.48, p < 0.0001$), followed by news media ($M = 4.10, t(2210) = 5.48, p < 0.0001$), infographic ($M = 2.57, t(2430) = 10.77, p < 0.0001$), and scientific visualizations ($M = 1.26, t(1947) = 10.56, p < 0.0001$). This implies that participants did more visual exploration between the data and other elements in the visualization from government sources. Examining the IEF by visualization type, there is no broad statistically significant trend with the exception of bar graphs which have the highest average IEF ($M = 5.01$) as compared to the second highest IEF visualization type of line graphs ($M = 3.99, t(1801) = 4.78, p < 0.00001$). This implies that participants did more visual exploration between the data and other elements with bar graphs. This may also relate to the observed trend with publication venue as the government publications have slightly higher numbers of bar and line visualizations as compared to the other sources.

Re-fixations: Of all the visual elements, across all visualization types and venus, legends have the highest re-fixation average ($M = 1.73$) followed by table header rows ($M = 1.49, t(2927) = 3.56, p < 0.0004$) and objects ($M = 1.38, t(6597) = 5.08, p < 0.0001$). Titles ($M = 1.27$) are the fourth most re-fixated element and significantly less than header rows ($t(4683) = 4.35, p < 0.0001$). The strong re-fixations of legends seems intuitive as one would expect viewing of a visualization with a legend to require a viewer to refer to the legend to interpret the data encoding. The same reasoning applies to use of a header row to interpret a table. More interesting is the significant re-fixations of the visualization's title (1-2 times during encoding) as this does not have the same function as a legend to enable interpretation of the data. However, the title may state the purpose or main point of the visualization thus serving as a reference and aid to the viewer. Finally, the elements with the lowest re-fixations include elements that you only need to look at once (e.g., paragraph text) or do not need to look at to understand the visualization (e.g., graphical element).

6.7.2 Memorability versus Recognizability

During the recognition phase of the study, HIT rates (HR) were generated for each visualization (i.e., what percentage of participants correctly identified the visualization as having seen it during the encoding phase). The mean recognizability HR is 83.30% ($SD = 17.72\%$), as compared to the previous study's memorability mean HR of 55.61% ($SD = 15.30\%$). The HR for recognizability, as well as the HR from the previous study on memorability (Chapter 5), for all the target visualizations are plotted in Fig. 6.6. For the recognizability HR, there are 121 visualizations ranked "1" in which every person correctly identified the image during recognition (top left portion of the plot), and there are no visualizations that have 0% recognition (lower right of the plot).

The increase in absolute HR values for recognizability as compared to memorability is to be expected since the visualizations were viewed longer during encoding

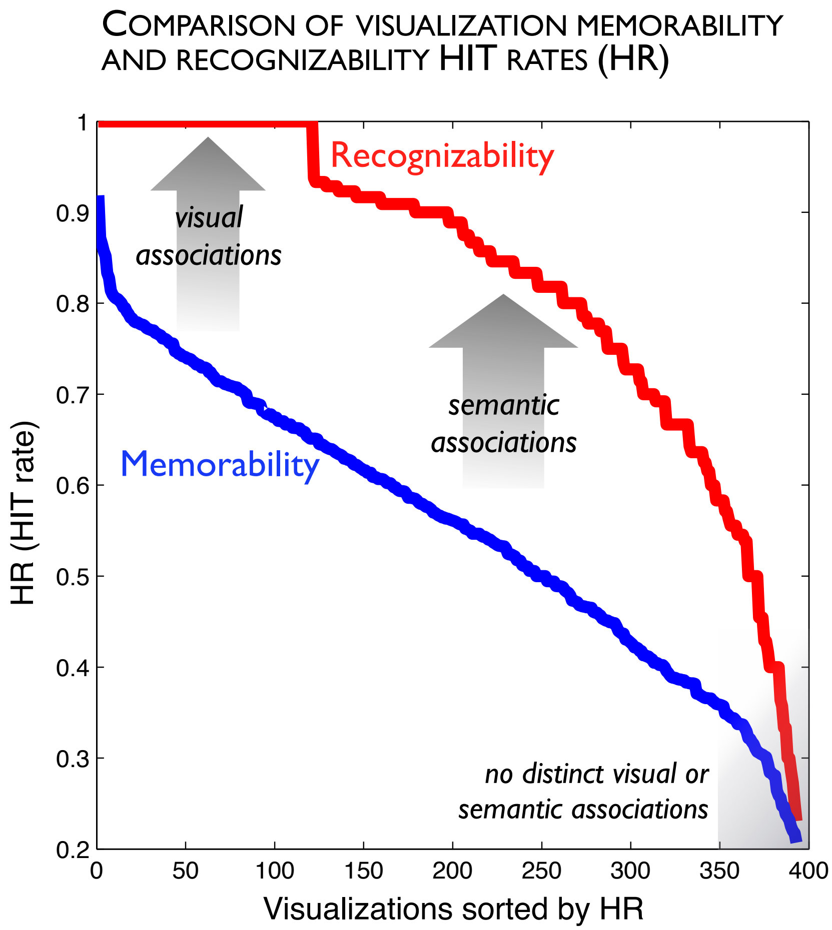

Figure 6.6: Plot comparing the recognizability HR (HIT rate) measures from the recognition phase of this experiment to the memorability HR (Chapter 5) of the target visualizations. The most memorable visualizations are still the most recognizable, most likely due to visual associations. The increase in recognizability for some less memorable visualizations is probably due to semantic associations (e.g., title, text, etc.).

(1 second versus 10 seconds). Also, if you compare the ranked order of visualizations, the most memorable are still the most recognizable and the least memorable are still the least recognizable. However, 13% of the middle-ranked memorability visualizations are now ranked more highly in recognizability. What elements of the visualizations make them more recognizable, and why do some visualizations remain forgettable even with longer encoding times?

We can gain insight into these questions by comparing the participants' fixations from the encoding phase and recognition phase of the experiment, as well as comparing the most and least recognizable visualization fixation patterns. As shown in Fig. 6.7, the heat maps overlaid on the visualizations represent the average of all of the participants' fixations on the visualization. The fixation patterns in the encoding phase demonstrate patterns of visual exploration (e.g., view graph, read title, look at legend, etc.), correlating with the trends described in Sec. 6.7.1. These visual exploration patterns are seen in both the most and least recognizable visualizations. However, there is a distinct difference between the fixation heat maps of the most and least recognizable visualizations in the recognition phase. The most recognizable visualizations have a fixation bias towards the center of the visualization. This indicates that a fixation near the center, while presumably also gathering information through peripheral vision, provides sufficient information to recognize the visualization. In contrast, the least recognizable visualizations have fixation heat maps that parallel the fixation patterns of visual exploration in the encoding phase. This indicates that the participants need to search the visualization for an *association* (i.e., a visual anchor) that will help with visualization recognition.

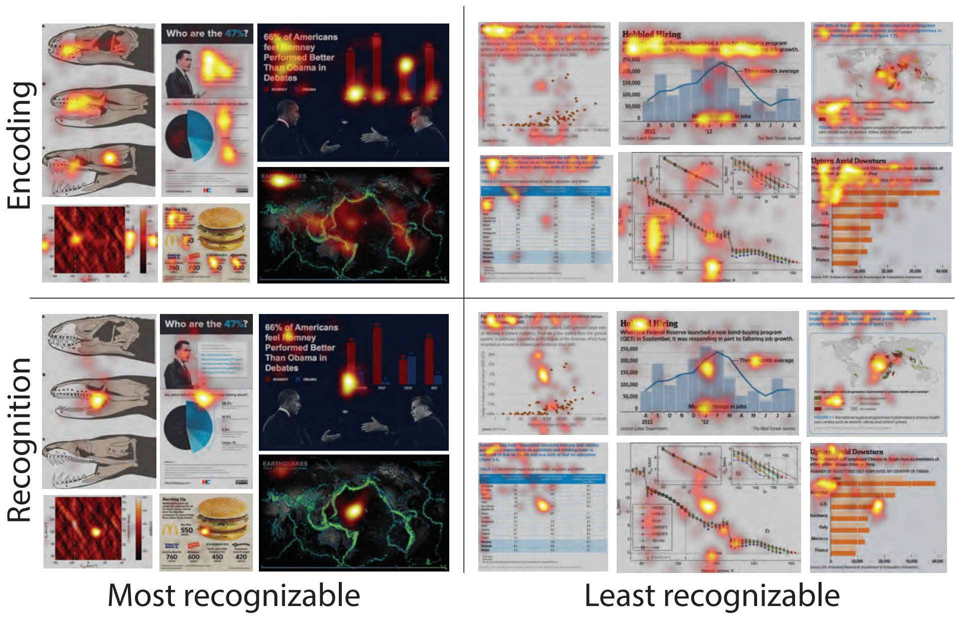

Figure 6.7: Examples of the most and least recognizable visualizations from our experiment. **TOP**: Eye-tracking fixation heat maps (i.e., average of all participants' gazes) from the *encoding* phase of the experiment in which each visualization was presented to the participant for 10 seconds. The fixation patterns demonstrate visual exploration of the visualization. **BOTTOM:** Eye-tracking fixation heat maps for the *recognition* phase of the experiment in which each visualization was presented for 2 seconds. The most recognizable visualizations all have a single focus in the center indicating quick recognition of the visualization, whereas the least recognizable visualizations have fixation patterns similar to the encoding fixations indicative of visual exploration (e.g., title, text, etc.) for recognition.

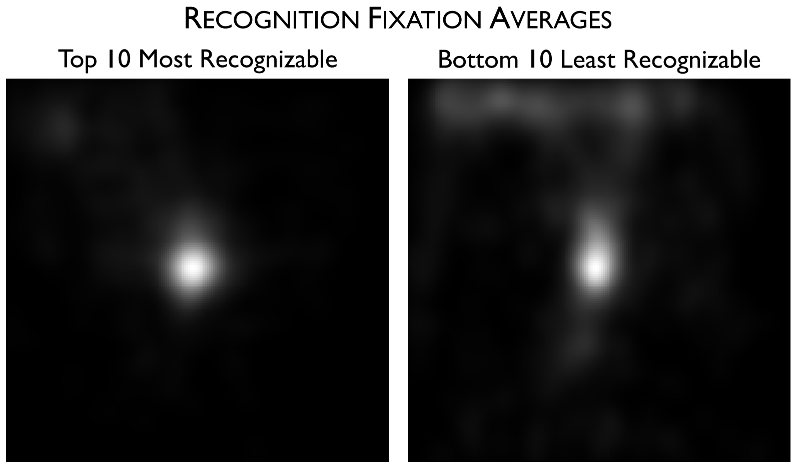

Figure 6.8: Average of the recognition fixation heat maps for the 10 most recognizable visualizations (left) and 10 least recognizable visualizations (right). The central focus on the left probably indicates recognition through visual association, whereas the smeared map on the right probably indicates visual search for semantic associations.

This contrast between the fixation patterns of the most and least recognizable visualizations is also distinct when the fixation heat maps are averaged for the most and least recognizable visualizations. As shown in Fig. 6.8, the most recognizable visualizations are characterized by a central focus. The least recognizable visualizations also have a central focus, but additionally have numerous other fixations in the visualization indicative of visual search for an association for recognition. The fixations along the top of the heat map for the least recognizable visualizations generally correspond to the location of the title and paragraph text describing the visualization in further detail. These fixations patterns indicate visual search for other associations for recognition.

In order to quantify and prove the significant difference between the fixation patterns of the most and least recognizable visualizations, we calculated four different metrics to quantify the difference in observed fixation patterns: the spatial variances in recognition fixation locations, the average recognition fixation distance viewed away from the image center, the Kullback-Leibler (KL) distance (defined below) between the center focus of the visualization and recognition fixations, and the KL distance between the encoding and recognition fixations.

In order to evaluate the variability of the recognition fixations, for each participant we calculated the variances of the x and y fixation locations over all of the recognition fixations. We then took the mean of these two variance values in order to produce an overall variance value. In order to determine if there is a difference in fixation variances across the participants for the most and least recognizable visualizations, we used the 121 visualizations that had a 100% recognizability HR and the 121 visualizations with the lowest recognizability HR. The most recognizable visualizations ($M = 19304.06$) have a significantly lower fixation variability as compared to the least recognizable visualizations ($M = 29103.54$) ($t(2710) = 19.59, p < 0.001$). Thus there is a significant difference in recognition fixation variability between the most and least recognizable visualizations with the most recognizable visualizations having less variance in the averaged fixations patterns.

Another measure to help distinguish between the most and least recognizable visualization recognition fixation patterns is to calculate the average distance from the visualization center at which participants fixate. This is calculated by taking the Euclidean distance between the center of the visualization and each of a participant's recognition fixations on the visualization, and computing the mean of all these distances. The mean distance of fixation from the center of the visualization is 182.18 pixels for the most recognizable visualizations and 248.11 pixels for the least recognizable visualizations ($t(2710) = 22.45, p < 0.001$). This demonstrates that participants' fixations on average look further away from the center in the least recognizable visualizations as compared to the most recognizable visualizations.

The last two measures which demonstrate that the fixation patterns during recognition are different for the most and least recognizable visualizations are based on the concept of KL distance. The KL distance (or divergence) is a distance metric which measures the difference between two continuous distributions, with a higher value indicative of a greater difference between the two distributions. In our first calculation, we compare the distance between a prior map representing the center of the visualization to the recognition fixation locations. The center prior map was created by convolving a single fixation located in the exact center of a visualization, with a Gaussian width (σ) of 25 pixels, which is approximately 1 degree of visual angle. This Gaussian width helps takes into account the noise during fixation measurement and helps make a continuous fixation map out of discrete fixations. A recognition fixation map was created for each participant by convolving all of their fixations on a single visualization with a Gaussian filter ($\sigma = 25$). A KL distance was computed between the center prior map and recognition fixation map for each participant and each visualization with all calculations averaged over all participants and all most or least recognizable visualizations. Laplacian smoothing was used to make all computations defined. The average KL distance between the center prior and recognition fixations is higher for the least recognizable visualizations ($M = 14.35$) as compared to the most recognizable visualizations ($M = 11.20$) ($t(2710) = 11.91, p < 0.001$). This means that there is a greater difference between the center prior and fixation patterns for the least recognizable visualizations, and that the fixation patterns for the most recognizable visualizations are closer to the central prior.

Finally, we calculate the KL distance between the encoding and recognition fixation locations. A recognition fixation map was created for each participant by convolving all of their fixations on a single visualization with a Gaussian filter ($\sigma = 25$). An encoding fixation map for the same participant and visualization was constructed analogously. The KL distance was then computed between the participant's encoding and recognition fixation maps for the same visualization, and the results were separately averaged across participants for the most and least recognizable visualizations. The average KL distance between the encoding and recognition fixations is higher for the most recognizable visualizations ($M = 14.87$) as compared to the least recognizable visualizations ($M = 13.80$) ($t(2710) = 5.83, p < 0.001$). This means that the encoding and recognition fixation patterns are more similar for the least recognizable visualizations demonstrating that the recognition fixations mimic the visualization exploration fixations of the encoding phase. Thus the participants are visually searching for associations to aid in recognition with the less recognizable visualizations. For the most recognizable visualizations the high KL distance is probably due to the recognition centrally-focused fixation being very different than the encoding fixations (i.e., more closely associated with the center prior map).

An additional calculation to quantify the difference between the most and least recognizable visualizations is to examine the average number of distinct foci in the recognition fixations. Qualitatively based on the fixation maps in Fig. 6.7 and Fig. 6.8, one would predict that the most recognizable visualizations would have $\sim$1 focus whereas the least recognizable visualizations would have multiple foci. For the 121 most recognizable and the 121 least recognizable visualizations, we applied thresholds to the recognition fixation maps ranging from 0.1-0.9 and counted the number of connected components. (Each fixation map is scaled so that the intensity values go from 0 to 1.) Qualitatively examining the fixation maps, a threshold of 0.4 appeared to be the most appropriate. This threshold measures an average of 1.45 connected components (i.e., distinct foci) in the most recognizable visualization fixations maps and an average of 2.94 connected components in the least recognizable visualizations ($t(240) = 8.92, p < 0.001$). (All threshold steps from 0.1 to 0.9 result in fewer foci in the most recognizable visualizations with statistical significance of $p < 0.001$.) Thus the least recognizable visualizations have more foci than the most recognizable visualizations indicative of visual search for an association for recognition.

Based on all of these different metrics, we see that **the recognition fixations are significantly different between the least recognizable visualizations and the most recognizable visualizations**, and that there are more distinct foci in the least recognizable visualizations. Thus there is more visual movement and exploration during recognition for the least recognizable visualizations. In the following sections we will discuss some possible explanations for this difference and what kinds of associations people use to recognize visualizations.

6.7.3 Visual versus Semantic associations

We have demonstrated that there is a distinct difference between the fixation patterns of the most and least recognizable visualizations. Which visual elements in the visualization contribute to this difference, and which visual elements help explain the overall HR increase for recognizability? We hypothesize that people are utilizing two different types of associations to help with visualization recognition: *visual associations* and *semantic associations*. The most recognizable visualizations consist of more visually distinct types of visualizations (i.e., diagrams, tree and network diagrams, and maps) as compared to the least recognizable visualizations. The most recognizable visualizations also contain a higher percentage of human recognizable objects (56%) as compared to the least recognizable visualizations (9%). Both distinct visualization types and human recognizable objects are examples of visual associations.

The visualizations that are not visually distinct are probably recognizable due to semantic associations. These semantic associations may include, for example, the visualization's title. This is observable in the visual elements with the highest total fixation time during the recognition portion of the experiment. The elements with the highest TFT during recognition are the titles ($M = 253.80ms$) and objects ($M = 252.73ms$) ($t(4944) = 4.19, p < 0.0001$). The titles serve as a semantic association, and the objects as a visual association for recognition. This trend of the title as a semantic association is also visually evident in Fig. 6.7 and Fig. 6.8 in which the fixation blurs outside of the central focus in the least recognizable visualizations during recognition correspond to the visualizations' title and other textual elements near the top. Thus these types of semantic associations are used for recognition if the visualization is not visually distinct or does not contain sufficient visual associations.

The visualizations that are least recognizable probably do not have a strong visual association nor a strong semantic association to assist with recognition (Fig. 6.6, lower right). The least recognizable visualizations, which have significant overlap and correlate with the least memorable visualizations from the previous study in Chapter 5, consist primarily of bar, line, point, and table visualizations. These types of visualizations were found to be the least memorable in the previous study. These visualization types are the least visually distinct. Also, most of the least recognizable visualizations come from government publications $M = 0.72, t(195) = 4.04, p < 0.0001$). Government visualizations tend to use the same templates and similar aesthetics, thus may contribute to confusion between visualizations during recognition, and the government visualizations are heavily composed of the least recognizable and least memorable visualization types including bar and line graphs.

6.7.4 Text Analysis

During the recall phase of the experiment, each participant provided text to answer the question "Describe the visualization in as much detail as possible" for visualizations they correctly recognized. In order to quantify these text descriptions, we manually coded every word of the description to mark whether the word was relevant to the visualization. In order to do this, we first recorded in our database the title of each visualization. Next we read through all of the participant generated text descriptions and manually coded whether each word is associated with a specific label in the visualization (Table 6.1), is verbatim text from the title, is irrelevant (e.g., articles such as "the", excess commentary like "I don't know", etc.), or is related to the information presented in the visualization but is not directly encoded in the visualization itself (i.e., participant used and synthesized prior knowledge of the topic in the text description).

For the 393 target visualizations, the study participants generated 2,063 text descriptions. The mean length of a description is 13.65 words ($SD = 10.13$), with a mean number of descriptions per visualization of 5.25 ($SD = 2.72$). A total of

28,152 words were manually coded from these text descriptions of which 18,698 words were deemed *relevant* to the visualization (i.e., these are the words that can be found directly in the visualization text or title, or directly mention a specific element of the visualization).

For all of the visualizations, the proportion of recall descriptions that contained at least one word that was relevant to the visualization was higher for the most recognizable visualizations ($M = 94\%$) as compared to the least recognizable visualizations ($M = 86\%$) ($t(229) = 3.50, p < 0.001$). Examining only the visualizations that have at least 3 recall text descriptions (327 out of the total 393 visualizations), the 100 most recognizable of these visualizations have text descriptions that contain on average 95% relevant words as compared to 87% relevant words in the 100 least recognizable visualization descriptions ($t(198) = 3.78, p < 0.001$). Thus on average the recall descriptions for the most recognizable visualizations contained more relevant words as compared to the least recognizable visualizations.

The length and content of these recall descriptions provide further insight and support for the difference between the most and least recognizable visualizations, as well as evidence to support the role of semantic associations to aid recognition. The average text description length is significantly shorter for the visualizations that are both least memorable and least recognizable ($M = 12.61$ *words*) as compared to the most memorable and most recognizable visualizations ($M = 15.05$ *words*, $t(284) = 3.58, p < 0.001$). This implies that participants recalled fewer details about the least recognizable visualizations.

We also observe shorter recall text descriptions for visualizations from government sources ($M = 11.43$ *words*) as compared to infographic visualizations which have the longest text descriptions ($M = 14.92\ words, t(156) = 4.12, p < 0.001$). Thus participants were able to recall more information about the infographic visualizations which, as previously discussed, are more recognizable than visualizations from government sources.

Looking at specific visual elements mentioned in the text descriptions, participants reference the text paragraph labels significantly more in the most recognizable visualizations as compared to the least recognizable visualizations. On average, for the most recognizable visualizations 33% of the recall text description is directly related to the paragraph text elements in the visualization and is significantly more than the 19% for the least recognizable visualizations ($t(125) = 2.46, p < 0.015$). This shows that for the most recognizable visualizations, the text and prose in the visualizations persisted in participants' memory through recall and may have contributed to recognition.

6.7.5 Summary of key observations

Based on the results presented in the proceeding sections, we summarize below the key observations:

"The more text in your visualization, the more time people will spend looking at it." One of the common observations throughout encoding is that if there is text on a visualization then people will spend time reading it. This was observed with the text titles, paragraphs, annotations, and legends during encoding with the TFT, FTA, and re-fixation measures. In the recall text descriptions, the most recognizable visualizations had significantly more words directly relating to the paragraph text in the visualizations as compared to the least recognizable visualizations. Thus not only do people spend time reading text during encoding, and use it for recognition, but it can help with visualization recall.

"Have a good title with your visualization." This trend is specifically observed during recognition as the element with the highest fixation time is the visualization's title. Titles also had a significant number of re-fixations during encoding. People not only notice the title, but refer back to it and probably use it as a semantic association for recognition. Thus having a good clear title that helps the viewer make an association with the visualization will help with recognition.

"People recognize a visualization either through visual associations or through semantic associations." The most recognizable visualizations have distinct visual characteristics, and their fixation patterns during recognition exhibit a central focus characteristic of recognition without visual search for semantic associations. The least recognizable visualizations have a significantly different fixation pattern with more distinct foci indicative of visual search for an association for recognition. The fixations during recognition focus on textual elements supporting the hypothesis of the use of semantic associations for recognition when a visual association is not distinct. Thus, if your visualization does not have a visually distinct association, it can still be recognizable with the inclusion of a strong supportive semantic association.

6.8 Conclusions & Future Work

Based on the results of this experiment, we identified the visual elements which contribute to visualization recognition and recall. We demonstrated the importance of text in a visualization both as a semantic association for recognition as well as a key component that people will spend significant time viewing and encoding. We also show that people recognize a visualization either through visual or semantic associations, and that the most recognizable visualizations are recognized with visual associations.

In future work, we hope to study subtle factors that may contribute to a visualization's recognizability. These factors include, for example, a participants emotional state (e.g., [60]) and their emotional response to a visualization (e.g., if it is about death, love, etc.). Another factor to evaluate and consider is the role of aesthetics in visualization design and how it may effect recognizability (e.g., [35, 139]). Finally, we hope to take advantage of other experimental set-ups past eye-tracking to understand how people respond to visualizations in the context of memorability and recognizability, such as EEG experiments (e.g., [3, 129]), as well as develop systems and apply vision models to understand the features a person sees with their peripheral vision not necessarily captured in the eye-tracking data. With a better understanding of visualization recognizability and recall, we have a more solid understanding of how people remember and recall visualizations and can more effectively design future studies to look at even higher level congnitive functions such as comprehension and engagement.

7

Conclusions & Future Work

Visualization is a powerful tool for data exploration and discovery. In order to understand what makes a visualization effective, and how to optimize the effectiveness for a given task, one needs to apply what we know about human perception and cognition. Through the application of evaluation methodologies we are able to more throughly understand these fundamental perceptual and cognitive principles and how they apply to visualization design, as well as develop new basic theory. This thesis presented two domain-specific case studies in the fields of biomedicine (Chapters 3) and computer science (Chapter 4) and two generalized experiments (Chapters 5 and 6) to better understand existing theory and develop novel fundamental theory.

In Chapter 3 we presented a new heart disease diagnostic tool, HemoVis, which utilizes a novel 2D tree diagram representation of arteries. Through an evaluation with real users and medical professionals we evaluated the effectiveness of the tool, 2D versus 3D representations, and color maps for identifying diseased regions. The results demonstrate the superior effectiveness of the 2D representation for a spatial-based task as well as inferiority of the rainbow color map. The 2D representation and non-rainbow color map resulted in more accurate and efficient diagnoses of the patient data.

In Chapter 4 we presented another case study in which we developed a new visualization tool, InProv, which utilizes a radial layout for the visualization of filesystem provenance data. We also developed a new time-based hierarchical node grouping method to present the data in a manner more intuitive to the user. We conducted an evaluation to compare InProv to Orbiter, a conventional tool which utilizes a node-link diagram representation. The results of our study demonstrated that for large complex datasets the radial layout of InProv and new time-based node grouping method were more accurate and efficient for completing tasks.

In Chapters 5 and 6 we presented the results of two different evaluations designed to understand fundamental memorability properties of visualizations. The first experiment presented in Chapters 5 was conducted on Amazon's Mechanical Turk in which we collected memorability scores for hundreds of visualizations from a variety of publication sources, and discovered that observers are consistent in which visualizations they find memorable and forgettable. We observed that some of the main factors that contribute to visualization memorability include human recognizable objects, color, unique visualization types, and overall visual distinctiveness.

Finally, in Chapter 6 we expanded on the results of Chapter 5 by exploring exactly which visual elements of a visualization contribute to memorability as well as visualization recognizability and recall through an eye-tracking laboratory evaluation. For the study we manually labeled all of the visual elements in hundreds of visualizations. During the experiment participants first viewed dozens of visualizations for 10 seconds each as part of an encoding phase, then saw these visualization mixed with other previously unseen visualization as part of a recognition task, and finally participants were asked to perform a recall task. The results of the experiment demonstrated the importance of text in a visualization and that people appear to recognize a visualization either through visual or semantic associations.

The observations and evaluation results on perceptual and cognitive principles presented in this thesis are broadly applicable across the field of visualization. In addition to the domain-specific broader impacts discussed in Sections 3.7 and 4.10, the results presented in Chapters 3 and 4 on visual encoding, color, spatial layout, complexity reduction, and dimensionality are all applicable to other areas of visualization involving spatial tasks. The results presented in Chapters 5 and 6 are fundamental theory thus are inherently broadly applicable to all areas of visualization design.

This area of research at the intersection of perception, cognition, and visualization theory is a very rich field of study with much potential. There are not enough evaluations as part of design studies across specific domain case studies, especially in the sciences, to evaluate or study perceptual and cognitive principles. These evaluations could collectively greatly expand our knowledge one evaluation at a time. These evaluations, such as those presented in Chapters 3 and 4, have the potential to both test if fundamental theory holds-up in the real world across domains, as well as create new theory.

Starting at the most low-level of visualization cognition theory, understanding how people remember and recall visualizations is also just the first step towards being able to understand visualization comprehension. Once we have a solid theory of perception and basic cognitive principles like memorability and recall, then we can move on as a community to studying higher level cognitive principles such as visualization aesthetics, engagement, and comprehension [24, 128]. Once we build up to these higher level functions then we will truly have a complete picture of how we as humans see, interpret, and understand visualizations, and how to design effective visual representations.

Bibliography

[1] James Abello, Frank van Ham, and Neeraj Krishnan. Ask-graphview: A large scale graph visualization system. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):669–676, 2006.

[2] Robert Amar, James Eagan, and John Stasko. Low-level components of analytic activity in information visualization. In *IEEE Symposium on Information Visualization*, pages 111–117, 2005.

[3] Erik W Anderson, Kristin C Potter, Laura E Matzen, Jason F Shepherd, Gilbert A Preston, and Cláudio T Silva. A user study of visualization effectiveness using eeg and cognitive load. *Computer Graphics Forum*, 30(3): 791–800, 2011.

[4] Richard C Anderson and James W Pichert. Recall of previously unrecallable information following a shift in perspective. *Journal of Verbal Learning and Verbal Behavior*, 17(1):1–12, 1978.

[5] Natalia Andrienko, Gennady Andrienko, and Peter Gatalsky. Exploratory spatio-temporal visualization: an analytical review. *Journal of Visual Languages & Computing*, 14(6):503 – 541, 2003.

[6] Daniel Archambault, Tamara Munzner, and David Auber. Grouse: Feature-based, steerable graph hierarchy exploration. In *Proceedings of the 9th Joint Eurographics/IEEE VGTC conference on Visualization*, pages 67–74. Eurographics Association, 2007.

[7] Daniel Archambault, Tamara Munzner, and David Auber. Grouseflocks: Steerable exploration of graph hierarchy space. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):900–913, 2008.

[8] Daniel Archambault, Tamara Munzner, and David Auber. Tuggraph: Path-preserving hierarchies for browsing proximity and paths in graphs. In *IEEE PacificVis '09*, pages 113 –120, April 2009.

[9] Wilma A Bainbridge, Phillip Isola, and Aude Oliva. The intrinsic memorability of face photographs. *Journal of Experimental Psychology: General*, 142 (4):1323–1334, 2013.

[10] Moshe Bar and Maital Neta. Humans prefer curved visual objects. *Psychological science*, 17(8):645–648, 2006.

[11] Albert-Laszlo Barabasi. The origin of bursts and heavy tails in human dynamics. *Nature*, 435(7039):207–211, 2005.

[12] Scott Bateman, Regan L Mandryk, Carl Gutwin, Aaron Genest, David McDine, and Christopher Brooks. Useful junk?: the effects of visual embellishment on comprehension and memorability of charts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2573–2582, 2010.

[13] Ben Bauer, Pierre Jolicoeur, and William B Cowan. Distractor heterogeneity versus linear separability in colour visual search. *Perception*, 25(11):1281 – 1293, 1996.

[14] Louis Bavoil, Steven P. Callahan, Patricia J. Crossno, Juliana Freire, and Huy T. Vo. Vistrails: Enabling interactive multiple-view visualizations. In *IEEE Visualization 2005*, pages 135–142, 2005.

[15] Laura Beckwith, Margaret Burnett, Susan Wiedenbeck, Curtis Cook, Shraddha Sorte, and Michelle Hastings. Effectiveness of end-user debugging software features: Are there gender issues? In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '05, pages 869–878, 2005.

[16] Laura Beckwith, Cory Kissinger, Margaret Burnett, Susan Wiedenbeck, Joseph Lawrance, Alan Blackwell, and Curtis Cook. Tinkering and gender in end-user programmers' debugging. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '06, pages 231–240, 2006.

[17] Christian Behrens. Infodesignpatterns. `http://www.niceone.org/infodesignpatterns/index.php5`, 2013.

[18] Jacques Bertin. *Graphics and graphic information processing*. Walter de Gruyter, 1981.

[19] Jacques Bertin. *Semiology of Graphics: Diagrams, Networks, Maps*. University of Wisconsin Press, 1983.

[20] Jacob Beutel, J Michael Fitzpatrick, Steven C Horii, Yongmin Kim, Harold L Kundel, Milan Sonka, and Richard L Van Metter. *Handbook of Medical Imaging, Volume 3. Display and PACS*. Washington, DC: SPIE Press, 2002.

[21] Hugh Beyer and Karen Holtzblatt. Contextual design. *interactions*, 6(1): 32–42, 1999.

[22] Olivier Biton, Sarah Cohen-Boulakia, Susan B. Davidson, and Carmem S. Hara. Querying and managing provenance through user views in scientific workflows. *IEEE International Conference on Data Engineering*, pages 1072–1080, 2008.

[23] Alan F Blackwell and TRG Green. Does metaphor increase visual language usability? In *IEEE Symposium on Visual Languages*, pages 246–253, 1999.

[24] Benjamin Samuel Bloom and David R Krathwohl. *Taxonomy of educational objectives: The classification of educational goals. Handbook I: Cognitive domain*. Longmans, 1956.

[25] Rita Borgo, Alfie Abdul-Rahman, Farhan Mohamed, Philip W Grant, Irene Reppa, Luciano Floridi, and Min Chen. An empirical study on using visual embellishments in visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2759–2768, 2012.

[26] David Borland and Russell M Taylor II. Rainbow color map (still) considered harmful. *IEEE Computer Graphics and Applications*, 27(2):14–17, 2007.

[27] Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

[28] Cynthia A Brewer. Color use guidelines for data representation. In *Proceedings of the Section on Statistical Graphics, American Statistical Association*, pages 55–60, 1999.

[29] Katja Bühler, Petr Felkel, Alexandra La Cruz, and Ra La Cruz. Geometric methods for vessel visualization and quantification - a survey. In *Geometric Modelling for Scientific Visualization*, pages 399–420. Springer-Verlag, 2002.

[30] Michael Burch, Natalia Konevtsova, Julian Heinrich, Markus Hoeferlin, and Daniel Weiskopf. Evaluation of traditional, orthogonal, and radial tree diagrams by an eye tracking study. *IEEE Transactions on Visualization and Computer Graphics*, 17(12):2440–2448, Dec 2011.

[31] Tor Busch. Gender differences in self-efficacy and attitudes toward computers. *Journal of Educational Computing Research*, 12(2):147–158, 1995.

[32] Alberto Cairo. *The Functional Art: An Introduction to Information Graphics and Visualization*. New Riders, 2013.

[33] Steven P. Callahan, Juliana Freire, Emanuele Santos, Carlos E. Scheidegger, Cláudio T. Silva, and Huy T. Vo. Vistrails: visualization meets data management. In *Proceedings of the 2006 ACM SIGMOD International Conference on Management of Data*, SIGMOD '06, pages 745–747, 2006.

[34] Sheelagh Carpendale. Evaluating information visualizations. In *Information Visualization*, volume 4950 of *Lecture Notes in Computer Science*, pages 19–45. Springer, 2008.

[35] Nick Cawthon and Andrew Vande Moere. The effect of aesthetic on the usability of data visualization. In *Information Visualization*, IV'07, pages 637–648, 2007.

[36] Yiannis S. Chatzizisis, Ahmet Umit Coskun, Michael Jonas, Elazer R. Edelman, Charles L. Feldman, and Peter H. Stone. Role of Endothelial Shear Stress in the Natural History of Coronary Atherosclerosis and Vascular Remodeling: Molecular, Cellular, and Vascular Behavior. *Journal of the American College of Cardiology*, 49(25):2379–2393, 2007.

[37] William S Cleveland and Robert McGill. Graphical perception: Theory, experimentation, and application to the development of graphical methods. *Journal of the American Statistical Association*, 79(387):531–554, 1984.

[38] Andy Cockburn and Bruce McKenzie. Evaluating the effectiveness of spatial memory in 2d and 3d physical and virtual environments. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 203–210, 2002.

[39] Mary Czerwinski, Desney S Tan, and George G Robertson. Women take a wider view. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '02, pages 195–202, 2002.

[40] Nicholas Del Rio and Paulo da Silva. Probe-it! visualization support for provenance. In *Advances in Visual Computing*, volume 4842 of *Lecture Notes in Computer Science*, pages 732–741. Springer Berlin / Heidelberg, 2007.

[41] G Draper, Yarden Livnat, and Richard F Riesenfeld. A survey of radial methods for information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 15(5):759–776, 2009.

[42] Yuri Englehardt. *The language of graphics: A framework for the analysis of syntax and meaning in maps, charts and diagrams*. PhD thesis, Institute for Logic, Language and Computation, University of Amsterdam, 2002.

[43] Gordon Erlebacher, David A Yuen, and Fabien Dubuffet. Current trends and demands in visualization in the geosciences. *Visual Geosciences formerly Electronic Geosciences*, 6(3):1–59, 2001.

[44] Stephen Few. The chartjunk debate: A close examination of recent findings. Visual Business Intelligence Newsletter, 2011.

[45] Stephen Few. Benefitting infovis with visual difficulties? Provocation without a cause. Visual Business Intelligence Newsletter, 2011.

[46] Aloke V. Finn, Masataka Nakano, Jagat Narula, Frank D. Kolodgie, and Renu Virmani. Concept of vulnerable/unstable plaque. *Arteriosclerosis, Thrombosis, and Vascular Biology*, 30(7):1282–1292, 2010.

[47] Andrew Forsberg, Jian Chen, and David H Laidlaw. Comparing 3d vector field visualization methods: A user study. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1219–1226, 2009.

[48] Andrew S. Forsberg, David H. Laidlaw, Andries van Dam, Robert M. Kirby, George E. Karniadakis, and Jonathan L. Elion. Immersive virtual reality for visualizing flow through an artery. In *Proceedings of the Conference on Visualization '00*, pages 457–460, 2000.

[49] Michael Friendly. A brief history of data visualization. In *Handbook of Data Visualization*, Springer Handbooks of Computational Statistics, pages 15–56. Springer Berlin Heidelberg, 2008.

[50] Yaniv Frishman and Ayellet Tal. Online dynamic graph drawing. *IEEE Transactions on Visualization and Computer Graphics*, 14(4):727–740, 2008.

[51] John Galloway and Simeon J. Simoff. Network data mining: methods and techniques for discovering deep linkage between attributes. In *Proceedings of the 3rd Asia-Pacific conference on Conceptual Modeling*, volume 53 of *APCCM '06*, pages 21–32, 2006.

[52] Emden R. Gansner, Yehuda Koren, and Stephen C. North. Topological fisheye views for visualizing large graphs. *IEEE Transactions on Visualization and Computer Graphics*, 11:457–468, 2005.

[53] Mohammad Ghoniem, J Fekete, and Philippe Castagliola. A comparison of the readability of graphs using node-link and matrix-based representations. In *IEEE Symposium on Information Visualization*, pages 17–24, 2004.

[54] Joseph H. Goldberg and Jonathan I. Helfman. Comparing information graphics: A critical look at eye tracking. In *Proceedings of the 3rd BELIV*

Workshop: BEyond Time and Errors: Novel evaLuation Methods for Information Visualization, BELIV '10, pages 71–78, 2010.

[55] Liang Gou and Xiaolong (Luke) Zhang. Treenetviz: Revealing patterns of networks over tree structures. *IEEE Transactions on Visualization and Computer Graphics*, 17(12), 2011.

[56] Steffen Hadlak, C Tominski, HJ Schulz, and H Schumann. Visualization of hierarchies in space and time. In *Workshop on Geospatial Visual Analytics: Focus on Time at the AGILE International Conference on Geographic Information Science*, 2010.

[57] Horst K. Hahn, Bernhard Preim, Dirk Selle, and Heinz Otto Peitgen. Visualization and interaction techniques for the exploration of vascular structures. In *Proceedings of the Conference on Visualization*, VIS '01, pages 395–402, 2001.

[58] Charles D Hansen and Chris R Johnson. *The Visualization Handbook*. Academic Press, 2005.

[59] Robert L Harris. *Information graphics: A comprehensive illustrated reference*. Oxford University Press, 1999.

[60] Lane Harrison, Drew Skau, Steven Franconeri, Aidong Lu, and Remco Chang. Influencing visual judgment through affective priming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2949–2958, 2013.

[61] Sandra G Hart. Nasa-task load index (nasa-tlx); 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 50(9):904–908, 2006.

[62] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Human mental workload*, 1(3):139–183, 1988.

[63] Amr Hassan and Christopher J Fluke. Scientific visualization in astronomy: Towards the petascale astronomy era. *Publications of the Astronomical Society of Australia*, 28(02):150–170, 2011.

[64] C.G Healey. Choosing effective colours for data visualization. *Proceedings of Visualization '96*, pages 263–270, 1996.

[65] Christopher G Healey. Perception in visualization. *Retrieved February 10 2008*, 2007.

[66] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: using mechanical turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 203–212, 2010.

[67] Jeffrey Heer and Ben Shneiderman. Interactive dynamics for visual analysis. *Queue*, 10(2):30, 2012.

[68] Jeffrey Heer, Michael Bostock, and Vadim Ogievetsky. A tour through the visualization zoo. *Communications of the ACM*, 53(6):59–67, 2010.

[69] Ivan Herman, Guy Melancon, and M. Scott Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000.

[70] Nigel Holmes. *Designer's Guide to Creating Charts and Diagrams*. Watson-Guptill Publications, 1984.

[71] Weidong Huang. Using eye tracking to investigate graph layout effects. In *International Asia-Pacific Symposium on Visualization*, APVIS '07, pages 97–100, 2007.

[72] Weidong Huang, P. Eades, and Seok-Hee Hong. A graph reading behavior: Geodesic-path tendency. In *IEEE Pacific Visualization Symposium*, PacificVis '09, pages 137–144, 2009.

[73] Jessica Hullman, Eytan Adar, and Priti Shah. Benefitting infovis with visual difficulties. *IEEE Transactions on Visualization and Computer Graphics*, 17 (12):2213–2222, 2011.

[74] David Huynh, David Karger, and Dennis Quan. Haystack: A platform for creating, organizing, and visualizing information using rdf. In *Semantic Web Workshop, Hawaii, USA*, 2002.

[75] Tobias Isenberg, Petra Isenberg, Jian Chen, Michael Sedlmair, and Torsten Moller. A systematic review on the practice of evaluating visualization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):2818–2827, 2013.

[76] Phillip Isola, Devi Parikh, Antonio Torralba, and Aude Oliva. Understanding the intrinsic memorability of images. In *Advances in Neural Information Processing Systems (NIPS)*, volume 24, pages 2429–2437, 2011.

[77] Phillip Isola, Jianxiong Xiao, Antonio Torralba, and Aude Oliva. What makes an image memorable? In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 145–152, 2011.

[78] Solid Source IT. Solid source explorer, 2009. URL `http://www.solidsourceit.com/`.

[79] T. J. Jankun-Kelly, Kwan-Liu Ma, and M. Gertz. A model and framework for visualization exploration. *IEEE Transactions on Visualization and Computer Graphics*, 13(2):357–369, 2007.

[80] Magesh Jayapandian, Adriane Chapman, V. Glenn Tarcea, Cong Yu, Aaron Elkiss, Angela Ianni, Bin Liu, Arnab Nandi, Carlos Santos, Philip Andrews, Brian Athey, David States, and H. V. Jagadish. Michigan molecular interactions (mimi): putting the jigsaw puzzle together. *Nucleic Acids Research*, 35(suppl 1):D566–D571, 2007.

[81] Chris Johnson. Top scientific visualization research problems. *IEEE Computer Graphics and Applications*, 24(4):13–17, 2004.

[82] Catherine M Jones and Susan D Healy. Differences in cue use and spatial memory in men and women. *Proceedings of the Royal Society B: Biological Sciences*, 273(1598):2241–2247, 2006.

[83] Armin Kanitsar, Dominik Fleischmann, Rainer Wegenkittl, Petr Felkel, and Meister Eduard Gröller. Cpr: curved planar reformation. In *Proceedings of the conference on Visualization '02*, VIS '02, pages 37–44, 2002.

[84] W Kapferer and T Riser. Visualization needs and techniques for astrophysical simulations. *New Journal of Physics*, 10(12):125008, 2008.

[85] Sung-Hee Kim, Zhihua Dong, Hanjun Xian, B. Upatising, and Ji Soo Yi. Does an eye tracker tell the truth about visualizations?: Findings while investigating visualizations for decision making. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2421–2430, 2012.

[86] Gordon Kindlmann, Erik Reinhard, and Sarah Creem. Face-based luminance matching for perceptual colormap generation. In *Proceedings of the conference on Visualization'02*, VIS '02, pages 299–306, 2002.

[87] Andreas Kjellin, Lars Winkler Pettersson, Stefan Seipel, and Mats Lind. Evaluating 2d and 3d visualizations of spatiotemporal information. *ACM Transactions on Applied Perception (TAP)*, 7(3):1–23, 2010.

[88] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *Journal of Experimental Psychology: General*, 139(3):558–578, 2010.

[89] Talia Konkle, Timothy F Brady, George A Alvarez, and Aude Oliva. Scene memory is more detailed than you think the role of categories in visual long-term memory. *Psychological Science*, 21(11):1551–1556, 2010.

[90] Christof Körner. Eye movements reveal distinct search and reasoning processes in comprehension of complex graphs. *Applied Cognitive Psychology*, 25(6):893–905, 2011.

[91] Robert Kosara, Christopher G. Healey, Victoria Interrante, David H. Laidlaw, and Colin Ware. User studies: Why, how, and when? *IEEE Computer Graphics and Applications*, 23(4):20–25, 2003.

[92] Stephen M Kosslyn. Understanding charts and graphs. *Applied Cognitive Psychology*, 3(3):185–225, 1989.

[93] Etien L Koua, Alan MacEachren, and M-J Kraak. Evaluating the usability of visualization methods in an exploratory geovisualization environment. *International Journal of Geographical Information Science*, 20(4):425–448, 2006.

[94] Martin I Krzywinski, Jacqueline E Schein, Inanc Birol, Joseph Connors, Randy Gascoyne, Doug Horsman, Steven J Jones, and Marco A Marra. Circos: An information aesthetic for comparative genomics. *Genome Research*, 19(9):1639–1645, 2009.

[95] Gautam Kumar and Michael Garland. Visual exploration of complex time-varying graphs. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):805–812, 2006.

[96] Markus Kunde, Henning Bergmeyer, and Andreas Schreiber. Provenance and annotation of data and processes. In Juliana Freire, David Koop, and Luc Moreau, editors, *IPAW '08*, chapter Requirements for a Provenance Visualization Component, pages 241–252. Springer-Verlag, 2008.

[97] A Kuß, M Gensel, B Meyer, VJ Dercksen, and S Prohaska. Effective techniques to visualize filament-surface relationships. *Computer Graphics Forum*, 29(3):1003–1012, 2010.

[98] David H Laidlaw, Robert M Kirby, Cullen D Jackson, J Scott Davidson, Timothy S Miller, Marco Da Silva, William H Warren, and Michael J Tarr. Comparing 2d vector field visualization methods: A user study. *IEEE Transactions on Visualization and Computer Graphics*, 11(1):59–70, 2005.

[99] Heidi Lam, Enrico Bertini, Petra Isenberg, Catherine Plaisant, and Sheelagh Carpendale. Empirical studies in information visualization: Seven scenarios. *IEEE Transactions on Visualization and Computer Graphics*, 18(9):1520–1536, 2012.

[100] Robert S Laramee, Helwig Hauser, Helmut Doleisch, Benjamin Vrolijk, Frits H Post, and Daniel Weiskopf. The state of the art in flow visualization: Dense and texture-based techniques. *Computer Graphics Forum*, 23 (2):203–221, 2004.

[101] Carol A Lawton. Gender, spatial abilities, and wayfinding. In *Handbook of Gender Research in Psychology*, pages 317–341. Springer, 2010.

[102] Bongshin Lee, Cynthia Sims Parr, Catherine Plaisant, Benjamin B Bederson, Vladislav Daniel Veksler, Wayne D Gray, and Christopher Kotfila. Treeplus: Interactive exploration of networks with enhanced tree layouts. *IEEE Transactions on Visualization and Computer Graphics*, 12(6):1414–1426, 2006.

[103] Edward Lee. Visualizing blog: A taxonomy of data visualization. `http://www.visualizing.org`, 2013.

[104] Ralph Lengler and Martin J Eppler. Color scales for image data. In *IASTED Proceedings of the Conference on Graphics and Visualization in Engineering*, 2007.

[105] Haim Levkowitz and Gabor T. Herman. Color scales for image data. *IEEE Computer Graphics and Applications*, 12(1):72–80, 1992.

[106] Adam Light and Patrick J Bartlein. The end of the rainbow? color schemes for improved data graphics. *EOS, Transactions American Geophysical Union*, 85(40):385–391, 2004.

[107] Dan R Lipşa, Robert S Laramee, Simon J Cox, Jonathan C Roberts, Rick Walker, Michelle A Borkin, and Hanspeter Pfister. Visualization for the physical sciences. *Computer Graphics Forum*, 31(8):2317–2347, 2012.

[108] Yarden Livnat, James Agutter, Shaun Moon, and Stefano Foresti. Visual correlation for situational awareness. In *IEEE Symposium on Information Visualization*, pages 95–102, 2005.

[109] Donald Lloyd-Jones and et al. Heart Disease and Stroke Statistics–2009 Update: A Report From the American Heart Association Statistics Committee and Stroke Statistics Subcommittee. *Circulation*, 119(3):e21–e181, 2009.

[110] Daisy Analysis Ltd. Daisy, 2003. URL `http://www.daisy.co.uk/`.

[111] Jock Mackinlay. Automating the design of graphical presentations of relational information. *ACM Transactions on Graphics*, 5(2):110–141, 1986.

[112] Peter Macko and Margo Seltzer. Provenance map orbiter: Interactive exploration of large provenance graphs. In *Proceedings of the 3rd Workshop on the Theory and Practice of Provenance (TaPP)*, 2011.

[113] Kim Marriott, Helen Purchase, Michael Wybrow, and Cagatay Goncu. Memorability of visual features in network diagrams. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2477–2485, 2012.

[114] Joel P Martin, J Edward Swan, Robert J Moorhead, Zhanping Liu, and Shangshu Cai. Results of a user study on 2d hurricane visualization. *Computer Graphics Forum*, 27(3):991–998, 2008.

[115] Winter Mason and Siddharth Suri. Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*, 44(1):1–23, 2012.

[116] E. Mcgrath. Methodology matters: Doing research in the behavioral and social sciences. In *Readings in Human-Computer Interaction: Toward the Year 2000 (2nd ed)*, pages 152–169, 1995.

[117] Tony McLoughlin, Robert S Laramee, Ronald Peikert, Frits H Post, and Min Chen. Over two decades of integration-based, geometric flow visualization. *Computer Graphics Forum*, 29(6):1807–1829, 2010.

[118] Simone Melchionna, Massimo Bernaschi, Sauro Succi, Efthimios Kaxiras, Frank J. Rybicki, Dimitris Mitsouras, Ahmet U. Coskun, and Charles L. Feldman. Hydrokinetic approach to large-scale cardiovascular blood flow. *Computer Physics Communications*, 181(3):462–472, 2010.

[119] Miriah Meyer, Tamara Munzner, and Hanspeter Pfister. Mizbee: A multiscale synteny browser. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):897–904, 2009.

[120] Miriah Meyer, Michael Sedlmair, P Samuel Quinan, and Tamara Munzner. The nested blocks and guidelines model. *Information Visualization*, 2013.

[121] Luc Moreau and et al. Special issue: The first provenance challenge. *Concurrency and Computation: Practice and Experience*, 20(5):409–418, 2008.

[122] Kenneth Moreland. Diverging color maps for scientific visualization. In *Advances in Visual Computing*, volume 5876 of *Lecture Notes in Computer Science*, pages 92–103. Springer Berlin Heidelberg, 2009.

[123] Kiran-Kumar Muniswamy-Reddy, David A Holland, Uri Braun, and Margo I Seltzer. Provenance-aware storage systems. In *USENIX Annual Technical Conference, General Track*, pages 43–56, 2006.

[124] Tamara Munzner. A nested model for visualization design and validation. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):921–928, 2009.

[125] Ken Museth, Torsten Müller, and Anders Ynnerman. Model-free surface visualization of vascular trees. *IEEE/Eurographics Symposium on Visualization*, 2008.

[126] Otto Neurath. *From hieroglyphics to Isotype: A visual autobiography*. Hyphen Press, London, 2010.

[127] Tatiana Pasternak, James W. Bisley, and David Calkins. *Visual Processing in the Primate Brain*, chapter 4. John Wiley & Sons, Inc., 2003.

[128] Robert E. Patterson, Leslie M. Blaha, Georges G. Grinstein, Kristen K. Liggett, David E. Kaveney, Kathleen C. Sheldon, Paul R. Havig, and Jason A. Moore. A human cognition framework for information visualization. *Computers & Graphics*, 2014.

[129] Evan M M Peck, Beste F Yuksel, Alvitta Ottley, Robert JK Jacob, and Remco Chang. Using fnirs brain sensing to evaluate information visualization interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 473–482, 2013.

[130] Zhenmin Peng and Robert S Laramee. Higher dimensional vector field visualization: A survey. In *Theory and Practice of Computer Graphics*, pages 149–163, 2009.

[131] Donna J. Peuquet. It's about time: A conceptual framework for the representation of temporal dynamics in geographic information systems. *Annals of the Association of American Geographers*, 84(3):441–461, 1994.

[132] Steven Pinker. A theory of graph comprehension. *Artificial intelligence and the future of testing*, pages 73–126, 1990.

[133] Catherine Plaisant. The challenge of information visualization evaluation. In *Proceedings of the ACM Working Conference on Advanced Visual Interfaces*, pages 109–116, 2004.

[134] Mathias Pohl, Markus Schmitt, and Stephan Diehl. Comparing the readability of graph layouts using eyetracking and task-oriented analysis. In

Proceedings of the Fifth Eurographics conference on Computational Aesthetics in Graphics, Visualization and Imaging, Computational Aesthetics '09, pages 49–56, 2009.

[135] Bernhard Preim and Steffen Oeltze. 3d visualization of vasculature: an overview. In *Visualization in medicine and life sciences*, pages 39–59. Springer, 2008.

[136] A. Johannes Pretorius and Jarke J. Van Wijk. Visual inspection of multivariate graphs. *Computer Graphics Forum*, 27(3):967–974, 2008.

[137] Michael Raschke, Tanja Blascheck, Marianne Richter, Tanja Agapkin, and Thomas Ertl. Visual analysis of perceptual and cognitive processes. In *Proceedings of the International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (IVAAP)*, 2014.

[138] Rolf Reber, Norbert Schwarz, and Piotr Winkielman. Processing fluency and aesthetic pleasure: is beauty in the perceiver's processing experience? *Personality and social psychology review*, 8(4):364–382, 2004.

[139] Katharina Reinecke, Tom Yeh, Luke Miratrix, Rahmatri Mardiko, Yuechen Zhao, Jenny Liu, and Krzysztof Z. Gajos. Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '13, pages 2049–2058, 2013.

[140] Penny Rheingans. Color, change, and control for quantitative data display. In *IEEE Conference on Visualization*, Visualization '92, pages 252–259, 1992.

[141] Penny L Rheingans. Task-based color scale design. In *28th AIPR Workshop: 3D Visualization for Data Exploration and Decision Making*, pages 35–43, 2000.

[142] Felix Ritter, Christian Hansen, Volker Dicken, Olaf Konrad, Bernhard Preim, and H-O Peitgen. Real-time illustration of vascular structures. *IEEE Transactions on Visualization and Computer Graphics*, 12(5):877–884, 2006.

[143] Richard A Robb. *Biomedical imaging, visualization, and analysis*. John Wiley & Sons, Inc., 1999.

[144] Jonathan C Roberts. Display models-ways to classify visual representations. *International Journal of Computer Integrated Design and Construction*, 2(4):241–250, 2000.

[145] Jose F Rodrigues, Agma JM Traina, Maria Cristina F de Oliveira, and C Traina. Reviewing data visualization: an analytical taxonomical study. In *IEEE International Conference on Information Visualization*, IV 2006, pages 713–720, 2006.

[146] B Rogowitz and Alan D Kalvin. The which blair project: A quick visual method for evaluating perceptual color maps. In *Proceedings of IEEE Visualization*, VIS'01, pages 183–190, 2001.

[147] Bernice Rogowitz and Lloyd A Treinish. Data visualization: the end of the rainbow. *IEEE Spectrum*, 35(12):52–59, 1998.

[148] Bernice E Rogowitz, Lloyd A Treinish, and Steve Bryson. How not to lie with visualization. *Computers in Physics*, 10(3):268–273, 1996.

[149] Bernice E Rogowitz, Alan D Kalvin, Adar Pelah, and Aron Cohen. Which trajectories through which perceptually uniform color spaces produce appropriate colors scales for interval data? In *Color and Imaging Conference*, pages 321–326. Society for Imaging Science and Technology, 1999.

[150] Timo Ropinski, Sven Hermann, Rainer Reich, Michael Schafers, and Klaus Hinrichs. Multimodal vessel visualization of mouse aorta pet/ct scans. *IEEE Transactions on Visualization and Computer Graphics*, 15(6):1515–1522, 2009.

[151] Joel Ross, Lilly Irani, M. Six Silberman, Andrew Zaldivar, and Bill Tomlinson. Who are the Crowdworkers? Shifting Demographics in Mechanical Turk. In *CHI'10 Extended Abstracts on Human Factors in Computing Systems*, pages 2863–2872, 2010.

[152] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008.

[153] Frank J. Rybicki, Simone Melchionna, Dimitris Mitsouras, Ahmet U. Coskun, Amanda G. Whitmore, Michael Steigner, Leelakrishna Nallamshetty, Fredrick G. Welt, Massimo Bernaschi, Michelle Borkin, Joy Sircar, Efthimios Kaxiras, Sauro Succi, Peter H. Stone, and Charles L. Feldman. Prediction of coronary artery plaque progression and potential rupture from 320-detector row prospectively ecg-gated single heart beat ct angiography: Lattice boltzmann evaluation of endothelial shear stress. *The International Journal of Cardiovascular Imaging*, 25(2):289–299, 2009.

[154] Gerard Salton, James Allan, Chris Buckley, and Amit Singhal. Automatic analysis, theme generation, and summarization of machine-readable texts. *Science*, 264(5164):1421–1426, 1994.

[155] Carlos Scheidegger, David Koop, Emanuele Santos, Huy Vo, Steven Callahan, Juliana Freire, and Cláudio Silva. Tackling the provenance challenge one layer at a time. *Concurrency and Computation: Practice and Experience*, 20(5):473–483, 2008.

[156] H Schulz. Treevis. net: A tree visualization reference. *IEEE Computer Graphics and Applications*, 31(6):11–15, 2011.

[157] Michael Sedlmair, Miriah Meyer, and Tamara Munzner. Design study methodology: Reflections from the trenches and the stacks. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2431–2440, 2012.

[158] Juliet Popper Shaffer. Multiple hypothesis testing. *Annual Review of Psychology*, 46(1):561–584, 1995.

[159] Ben Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Symposium on Visual Languages*, pages 336–343, 1996.

[160] Ben Shneiderman. Why not make interfaces better than 3d reality? *IEEE Computer Graphics and Applications*, 23(6):12–15, 2003.

[161] Harri Siirtola, Tuuli Laivo, Tomi Heimonen, and K-J Raiha. Visual perception of parallel coordinate visualizations. In *IEEE International Conference on Information Visualisation*, pages 3–9, 2009.

[162] Yogesh Simmhan, Paul Groth, and Luc Moreau. Special section: The third provenance challenge on using the open provenance model for interoperability. *Future Generation Computer Systems*, 27(6):737–742, 2011.

[163] Peter Howard Stone, Shigeru Saito, Saeko Takahashi, Yasuhiro Makita, Shigeru Nakamura, Tomohiro Kawasaki, Akihiko Takahashi, Takaaki Katsuki, Sunao Nakamura, Atsuo Namiki, Atsushi Hirohata, Toshiyuki Matsumura, Seiji Yamazaki, Hiroyoshi Yokoi, Shinji Tanaka, Satoru Ohtsuji, Fuminobu Yoshimachi, Junko Honye, Dawn Harwood, Michail Papafaklis, Martha Reitman, Ahmet U. Coskun, and Charles L. Feldman. The prediction trial: In-vivo assessment of coronary endothelial shear stress, arterial remodeling, and plaque morphology to predict coronary atherosclerosis progression and rupture in man. In *Journal of the American College of Cardiology*, 2011.

[164] Matus Straka, Michal Cervenansky, Alexandra La Cruz, Arnold Kochl, Milos Sramek, Eduard Groller, and Dominik Fleischmann. The vesselglyph: Focus & context visualization in ct-angiography. In *IEEE Proceedings of the Conference on Visualization*, VIS '04, pages 385–392, 2004.

[165] Desney S Tan, Mary Czerwinski, and George Robertson. Women go with the (optical) flow. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 209–215, 2003.

[166] Maurice Termeer, Javier Oliván Bescós, Marcel Breeuwer, Anna Vilanova, Frans Gerritsen, and M Eduard Groller. Covicad: Comprehensive visualization of coronary artery disease. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1632–1639, 2007.

[167] Christian Tominski, James Abello, and Heidrun Schumann. Axes-based visualizations with radial layouts. In *Proceedings of the 2004 ACM Symposium on Applied Computing*, pages 1242–1247, 2004.

[168] Christian Tominski, Georg Fuchs, and Heidrun Schumann. Task-driven color coding. *IEEE International Conference on Information Visualisation*, pages 373–380, 2008.

[169] Melanie Tory and Torsten Moller. Rethinking visualization: A high-level taxonomy. In *IEEE Symposium on Information Visualization*, pages 151–158, 2004.

[170] Melanie Tory and Torsten Moller. Human factors in visualization research. *IEEE Transactions on Visualization and Computer Graphics*, 10(1):72–84, 2004.

[171] Melanie Tory, Arthur E Kirkpatrick, M Stella Atkins, and Torsten Moller. Visualization task performance with 2d, 3d, and combination displays. *IEEE Transactions on Visualization and Computer Graphics*, 12(1):2–13, 2006.

[172] Melanie Tory, David W Sprague, Fuqu Wu, Wing Yan So, and Tamara Munzner. Spatialization design: Comparing points and landscapes. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1262–1269, 2007.

[173] Edward Tufte. *The Visual Display of Quantitative Information*. Cheshire (Conn.), 1983.

[174] Edward Tufte. *Envisioning Information*. Cheshire (Conn.), 1990.

[175] Edward R Tufte and David Robins. *Visual explanations*, volume 25. Graphics Press Cheshire, CT, 1997.

[176] Sherry Turkle. Computational reticence: Why women fear the intimate machine. *Technology and women's voices: Keeping in touch*, pages 41–61, 1988.

[177] Mario Valle. Visualization support to chemistry research (poster). In *Proceeding of IEEE Visualization*, 2006.

[178] Stijn van Dongen. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.

[179] Andrew Vande Moere, Martin Tomitsch, Christoph Wimmer, Boesch Christoph, and Thomas Grechenig. Evaluating the effect of style in information visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2739–2748, 2012.

[180] Tatiana Von Landesberger, Arjan Kuijper, Tobias Schreck, Jörn Kohlhammer, Jarke J van Wijk, J-D Fekete, and Dieter W Fellner. Visual analysis of large graphs: State-of-the-art and future research challenges. *Computer Graphics Forum*, 30(6):1719–1749, 2011.

[181] Colin Ware. Color sequences for univariate maps: Theory, experiments and principles. *IEEE Computer Graphics and Applications*, 8(5):41–49, 1988.

[182] Colin Ware. Designing with a 2 1/2-d attitude. *Information Design Journal*, 10(3):258–65, 2001.

[183] Colin Ware. *Visual Thinking for Design*. Morgan Kaufmann, 2008.

[184] Colin Ware. *Information visualization: perception for design*. Elsevier, 2012.

[185] Lei Zhu, Steven Haker, and Allen R Tannenbaum. Conformal flattening maps for the visualization of vessels. *SPIE Medical Imaging*, 4681:742–748, 2002.

[186] Lei Zhu, Steven Haker, and Allen Tannenbaum. Flattening maps for the visualization of multibranched vessels. *IEEE Transactions on Medical Imaging*, 24(2):191–198, 2005.